2004

Genetic diversity among progenitors and derived lines of two maize (Zea mays L.) populations

James R. Rouse
Iowa State University

NOTE TO USERS

Genetic diversity among progenitors and derived lines of two maize (*Zea mays* L.) populations

by

James R. Rouse

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Plant Breeding

Program of Study Committee:
Kendall Lamkey, Major Professor
E. Charles Brummer
Jean-Luc Jannink
Mike Lee
John Nason

Iowa State University

Ames, Iowa

2004

Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

James R. Rouse

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program

TABLE OF CONTENTS

CHAPTER 1: GENERAL INTRODUCTION

This dissertation begins with a general introduction and a literature review, followed by two manuscripts to be submitted to *Crop Science*, and ends with a section of general conclusions. References of citations made within a chapter will be listed at the end of each chapter. An appendix will be used to convey some of the detailed information regarding the reciprocal recurrent selection program described below and the markers used in this study.

Brief History Of The Germplasm Used In This Study

Reciprocal recurrent selection (RRS), a method first proposed by Comstock et al. (1949), is intended to be used for simultaneous improvement of two germplasm sources. Selection criteria and selection intensity may vary, but the nucleus of the strategy involves using each population as a tester for the other. When using RRS in maize breeding programs, advanced cycles of selection should show improvements in both general and specific combining ability (Keeratinijakal and Lamkey, 1993). Shortly after the 1949 publication by Comstock et al., an RRS project was established at Iowa's Cooperative Federal-State Maize Breeding Program using the cycle 0 (C0) of two recently developed synthetic populations (Penny and Eberhart, 1971).

The maize populations Iowa Corn Borer Synthetic #1 (Corn Borer, or BSCB1) and Iowa Stiff Stalk Synthetic (Stiff Stalk, or BSSS) were produced by G.F. Sprague in the 1930s and 1940s. BSSS was originally created at the United States Department of Agriculture (USDA) farm in Arlington Farms, Virginia, although this was a cooperative project between Sprague in Virginia and M.T. Jenkins at Iowa State University (Penny, 1968). The 16 inbred lines used to form the BSSS population were chosen by various prominent breeders as those

that were "stiffest stalked" (Sprague, 1946). Furthermore, two of the 16 inbreds were recycled inbreds, so a total of 20 parent inbreds were involved (Troyer, 2004).

BSCB1 was formed in a similar manner to BSSS, but this population included only 12 inbred progenitors intermated in various crossing combinations. The detailed information surrounding the origin of BSCB1 is not as well documented as for BSSS. It is likely, though, that Sprague formed this population around the same time as BSSS, and brought it with him from Virginia to Iowa by way of Columbia, Missouri (O. Smith, personal communication; Hallauer, 1984). The selection of BSCB1 to be used in a reciprocal recurrent selection program with BSSS may have largely been due to Sprague's familiarity with the populations, since he was instrumental in their origination.

The RRS program initiated with C0 populations in 1949 is currently in the 16th cycle of selection—in 2004, C16 testcross progenies were made for evaluation in 2005. Iowa Stiff Stalk Synthetic is one of the most important source populations for lines with above average stalk quality and combining ability (Hallauer, 1984). During the 50-plus years of the RRS program at Iowa State, approximately 50 inbreds and 30 populations containing germplasm derived in whole or in part from this program have been released to the public (Rouse et al., 2003). Approximately two-thirds of the lines and populations are related to BSSS, with the remainder related to either BSCB1 or both source populations. At one time it was estimated that about 19% of the total hybrid seed needed to plant the 1980 U.S. maize acreage was made from inbred lines derived out of BSSS (Zuber and Darrah, 1981). This was considered a minimum estimate, however, because it did not include related inbred lines or proprietary lines derived from BSSS.

The longevity of the RRS program between BSSS and BSCB1, and the popularity of the germplasm developed from the program, has led to numerous studies involving BSSS and BSCB1. Details of the RRS program and its effects on various genetic and phenotypic parameters of the two populations can be found in Penny and Eberhart (1971); Eberhart et al. (1973); Martin and Hallauer (1980); Smith (1983); Lamkey et al. (1991); Keeratinijakal and Lamkey (1993); Schnicker and Lamkey (1993); and Holthaus and Lamkey (1995). There have also been numerous studies to examine the genetic structure, diversity, and other molecular genetic parameters of these populations. For a more thorough review, see Messmer et al. (1991); Labate et al. (1997; 1999; 2000); Hagdorn et al. (2003); Guimarães (2001); and Hinze (2003).

Notes About the Current Study

The study described in this dissertation is related to those described in the dissertation of Hinze (2003), and in Hinze et al. (in review). Her studies involved 85 SSR markers on the progenitor lines of BSSS and BSCB1, as well as C0, C1, C3, C6, C9, C12, and C15 of both populations. Hinze's main objective was to determine how the genetic structure of the populations has changed over time. The study described here is of a similar nature, but examines the genetic diversity of inbred lines derived from the populations, rather than the populations per se. Briefly, there are 227 SSR markers on the progenitors of BSSS and BSCB1, and on several inbred lines derived from these populations and the RRS program. The specific objectives were: To examine measures of genetic diversity between the progenitors and the derived lines; and to evaluate the genetic distance between the progenitors and the derived lines, considering the derived lines both as individuals and as

groups. We are using the inbred progenitors of, and derived inbred lines from, BSSS and BSCB1 to see if we can detect significant information from the changes in allele frequencies and gene diversity between the progenitor groups and their respective derived lines. The information obtained from the derived lines may lead us to conclusions about which marker loci, and possibly which alleles, may have undergone changes due to forces other than random genetic drift. The key is that we are using marker information from inbred lines, not the populations per se. Even though our lines are related to and derived from the BSSS and BSCB1 populations, we are not trying to make claims about the genetic makeup of the populations.

References

Comstock, R.E., H.F. Robinson, and P.H. Harvey. 1949. A breeding procedure designed to make maximum use of both general and specific combining ability. Agronomy Journal. 41:360-367.

Eberhart, S.A., S. Debela, and A.R. Hallauer. 1973. Reciprocal recurrent selection in BSSS and BSCB1 maize populations and half-sib selection in BSSS. Crop Sci. 13:451-456.

Guimarães, P. 2001. Effects of a long-term recurrent selection program on the genetic structure of the BSSS maize population. PhD dissertation. Iowa State University, Ames.

Hagdorn, S., K.R. Lamkey, M. Frisch, P.E.O. Guimarães, and A.E. Melchinger. 2003. Molecular genetic diversity among progenitors and derived elite lines of BSSS and BSCB1 maize populations. Crop Sci. 43:474-482.

Hallauer, A.R. 1984. George F. Sprague: 60 years of corn research. Maydica. 29:351-355.

Hinze, L.L. 2003. Molecular marker analysis of population genetic structure and progress from reciprocal recurrent selection in two Iowa maize (*Zea mays* L.) populations. PhD dissertation. Iowa State University, Ames.

Hinze, L.L., S. Kresovich, J. Nason, and K.R. Lamkey. (In review.) Population genetic diversity in a maize reciprocal recurrent selection program. Crop Sci.

Holthaus, J.F., and K.R. Lamkey. 1995. Response to selection and changes in genetic parameters for 13 plant and ear traits in two maize recurrent selection programs. Maydica. 40:357-370.

Keeratinijakal, V., and K.R. Lamkey. 1993. Responses to reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 33:73-77.

Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 1997. Molecular genetic diversity after reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 37:416-423.

Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 1999. Temporal changes in allele frequencies in two reciprocally selected maize populations. Theor. Appl. Genet. 99:1166-1178.

Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 2000. Hardy-Weinberg and linkage equilibrium estimates in the BSSS and BSCB1 random mated populations. Maydica. 45:243-255.

Lamkey, K.R., P.A. Peterson, and A.R. Hallauer. 1991. Frequency of the transposable element *Uq* in Iowa stiff stalk synthetic maize populations. Genet. Res. 114: 2561-2568.

Martin, J.M., and A.R. Hallauer. 1980. Seven cycles of reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 20:599-603.

Messmer, M.M., A.E. Melchinger, M. Lee, W.L. Woodman, E.A. Lee, and K.R. Lamkey. 1991. Genetic diversity among progenitors and elite lines from the Iowa Stiff Stalk Synthetic (BSSS) maize population: comparison of allozyme and RFLP data. Theor. Appl. Genet. 83:97-107.

Penny, L.H. 1968. Selection-induced differences among strains of a synthetic variety of maize. Crop Sci. 8:167-168.

Penny, L.H., and S.A. Eberhart. 1971. Twenty years of reciprocal recurrent selection with two synthetic varieties of maize (*Zea mays* L.). Crop Sci. 11:900-903.

Rouse, J.R., L.L. Hinze, and K.R. Lamkey. 2003. Selection history and developments from the BSSS and BSCB1 maize populations. *In* International Symposium on Plant Breeding poster presentations [CD-ROM]. CIMMYT, Mexico City, Mexico.

Schnicker, B.J., and K.R. Lamkey. 1993. Interpopulation genetic variance after reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 33:90-95.

Smith, O.S. 1983. Evaluation of recurrent selection in BSSS, BSCB1, and BS13 maize populations. Crop Sci. 23:35-40.

Sprague, G.F. 1946. Early testing of inbred lines of corn. J. Am. Soc. Agron. 38:108-117.

Troyer, A.F. 2004. Background of U.S. hybrid corn II: Breeding, climate, and food. Crop Sci. 44:370-380.

Zuber, M.S., and L.L. Darrah. 1981. 1979 U.S. corn germplasm base. Proc. 35th Ann. Corn & Sorghum Res. Conf. ASTA. pp. 234-249.

CHAPTER 2: LITERATURE REVIEW

There are several recent molecular genetic studies involving the Iowa Corn Borer Synthetic #1 (BSCB1) and Iowa Stiff Stalk Synthetic (BSSS) populations. For markers, some of the studies used restriction fragment length polymorphisms (RFLPs) and others used simple sequence repeats (SSRs). For germplasm, they used various combinations of the inbred progenitors of the populations, inbreds derived from the populations, and various cycles of the populations per se. This literature review discusses the findings of the most recent molecular genetic studies of the BSCB1 and BSSS, including some that are still being prepared for publication. Other studies whose methods or results may have had an impact on one or more of the studies involving these two populations are also cited.

The Germplasm

G.F. Sprague initiated a reciprocal recurrent selection (RRS) program in 1949 using the BSSS and BSCB1 cycle 0 (C0) populations (Penny and Eberhart, 1971). There were 16 inbred lines used to form the BSSS population (Sprague, 1946). Furthermore, two of the 16 inbreds were recycled inbreds, so a total of 20 parent inbreds were involved (Troyer, 2004). Most of the inbred progenitors have been maintained over time, but two of them (*F1B1* and *CI187*) have been lost. However the parental lines (*IndB2* and *Fe*) of one of the lost inbreds (*F1B1*) are still maintained, leading to 16 available BSSS progenitor inbreds (Labate et al., 1997). BSCB1 was formed from 12 inbred progenitors, all of which are still available. The appendix contains a full list of the inbreds used and the mating designs for both populations.

The RRS program that began in 1949 is currently in the 16th cycle of selection. In 2004, C16 testcross progenies were made and will be evaluated in 2005. During the 50-plus

years of the RRS program at Iowa State, approximately 50 inbreds and 30 populations containing germplasm derived in whole or in part from this program have been released to the public (Rouse et al., 2003). Approximately two-thirds of the lines and populations are related to BSSS, the remainder related to either BSCB1 or both source populations. It was estimated that about 19% of the total hybrid seed needed to plant the 1980 U.S. maize acreage was made from inbred lines derived out of BSSS (Zuber and Darrah, 1981). This was considered a minimum estimate, however, because it did not include related inbred lines or proprietary lines derived from BSSS.

BSSS has been a popular germplasm source not only for the RRS program, but several other recurrent selection programs as well. Because of this popularity, many inbreds have been developed that are derived from BSSS, but may not have been a product of the RRS program with BSCB1. In addition, many of the inbred lines were then selected, based on various favorable traits, to be among the progenitors of even more populations. For example: BSSS2, a population released in 1971, was derived from two strains of BSSS—one was developed after four cycles of RRS, and one was developed after six cycles of recurrent selection for general combining ability (Russell et al., 1971). Also, *B10*, *B14*, *B37*, *B40*, *B43*, and *B44*, all derived from BSSS, were among the progenitors of the population BSSSS—the Iowa Super Stiff Stalk Synthetic (Russell et al., 1971).

Genetic Analyses of Maize Breeding Populations

In the first published results of a molecular genetic study with BSSS, Neuhausen (1989) used RFLP markers on the BSSS progenitors, several inbreds released from BSSS cycles, some inbreds derived from recycled BSSS inbreds, and inbreds used as testers or

parents of testers for BSSS and BSSS-derived inbreds. The population used in this study, BSSS(HT), was undergoing half-sib inbred tester recurrent selection using *IA13* (a four-parent double cross hybrid) as the tester. Neuhausen examined 402 alleles at 70 loci, and found 22.6% unique alleles, or alleles that were represented by only one inbred among the BSSS progenitors. Only 28 of the 91 unique alleles were found in subsequent BSSS(HT) cycles or derived inbreds, and only nine of those alleles were found in the three inbreds that were released from C5, C6, and C7. Of the nine recovered alleles, four progenitors, *CI540*, *LE23*, *Ill. Hy*, *OS420*, contributed two alleles each, and *Ind. B2* contributed a single allele (Neuhausen, 1989).

Messmer et al. (1991) evaluated RFLP and allozyme data using the 16 available BSSS progenitors, four inbreds derived from the BSSS(HT) recurrent selection program, and the inbred *Mo17* (developed from BSSS progenitor *CI187* and *C103*). They found that genetic variation was substantially greater for RFLPs than for allozymes, but the proportion of unique alleles was similar—25% for the RFLPs and 27% for the allozymes. In their study, 17% of the unique alleles of both allozymes (two out of 12) and RFLPs (20 out of 117) were detected in the four elite lines used in the study. Four of the unique RLFP variants came from *Ind. TR9-1-1-6*, and three were from *Ind. 461-3*.

Early molecular studies involving the RRS program between BSSS and BSCB1 were performed by Labate et al. (1997; 1999). In the initial publication, Labate et al. (1997) described genetic diversity measures among the progenitors of the populations and C0 and C12 of the populations per se. As expected, the progenitors were highly homozygous, and the authors reported that no single progenitor made excessive genetic contributions to C0 or C12. Similar to previous reports, about 25% of the alleles among the progenitors were unique to a

single inbred. This held true for both BSSS and BSCB1. However, by C12 approximately 60% of the unique alleles had disappeared in both populations. In addition, the average number of alleles per locus decreased by approximately 33%, from about four to less than three.

A measure of genetic distance between the two progenitor groups showed that they were initially very closely related, with Nei's (1978) unbiased genetic distance of 0.07 (Labate et al., 1997). The distance between the BSSS and BSCB1 C12 populations was 0.66, indicating substantial divergence from the progenitors and the C0 populations.

Following their 1997 publication, Labate et al. published reports detailing the temporal changes in allele frequencies in the BSSS and BSCB1 populations (Labate et al., 1999), and estimates of Hardy-Weinberg and linkage equilibrium (Labate et al., 2000). In the study of temporal changes in allele frequency, the authors used plan II of Waples' (1989) test to determine if forces other than random genetic drift (i.e., selection) was the cause of the observed changes in allele frequency. About 17% of the observed loci rejected the null hypothesis that drift alone was responsible for their frequency changes between C0 and C12. These loci were evenly distributed throughout the genome, and appeared to fit a pattern of fixing complementary alleles, since none of these alleles were shared between the two populations (Labate et al., 1999).

Labate et al. (1999) also mentioned that the loci that rejected the null hypothesis may be hitchhiking loci, rather than selected loci, and thought that a more powerful test might identify more loci that rejected the null hypothesis, and perhaps a larger fraction would be shared. But even if that did occur, they point out that natural selection cannot be ruled out as influencing allele frequency changes in the populations.

The BSSS and BSCB1 populations are maintained using methods that intend to ensure random mating within the populations. If successful, both populations will show Hardy-Weinberg and linkage equilibrium at all loci. Labate et al. (2000) examined these parameters in the populations to determine if the observed genetic properties matched the theoretical expectations. Most of the loci examined in this study met the expectations of equilibrium, but there were a few that showed excess homozygosity or heterozygosity. The majority of loci that deviated from equilibrium showed excess homozygosity. In the article by Labate et al. (2000), and several others cited therein, the excess homozygosity was theorized to be due to problems with sample sizes or nonrandom mating. The homozygosity found in the BSSS and BSCB1 populations was likely due to positive assortative mating (Labate et al., 2000).

In 2001, Guimarães (Ph.D. dissertation) used RFLPs to assess changes in allele and genotype frequencies in a BSSS population that had undergone seven cycles of half-sib selection [this is BSSS(HT), in which *IA13* was used as the tester], followed by seven cycles of S_2 selection. The S_2 selection occurred in the populations BS13(S), which was formed from BSSS(HT)C7. In 1969, 1000 plants of BSSS(HT)C7 were selfed, and 288 ears were selected for prolificacy, ear height, stalk rot resistance, and early maturity. In 1970 the 288 S_1 lines were evaluated for cold tolerance and European corn borer [*Ostrinia nubilalis* (Hübner)] resistance, and 29 lines were selected to form the BS13(S)C0 (Lamkey, 1992). Therefore BSSS(HT)C7 is not identical to BS13(S)C0, and the three germplasm groups included in the study are the BSSS progenitors, BS13(S)C0, and BS13(S)C7. While the recurrent selection methods used were different than the RRS program using BSSS and

BSCB1, the BSSS progenitors and the initial C0 source population in this study are the same as in the RRS program.

Some of the analyses used were similar to those of Labate et al. (1997; 1999; 2000). Guimarães (2001) also examined temporal changes in RFLP genotypes and alleles and tested deviations for neutrality. As expected, the BSSS progenitors showed high variability for the alleles at each locus. Guimarães reported 30.9% of the 456 alleles detected in the progenitors were unique to a single inbred, a slight increase over the previously published results. The author also noted a significant reduction in the sampled genetic variability over the cycles of selection. The variability detected in the BS13(S)C0 was very similar to BS13(S)C7, rather than intermediate between the progenitors and the C7 (recall that BS13(S)C0 was formed after seven cycles of half-sib recurrent selection). The genetic distance between the BS13(S)C0 and C7 indicated the populations were closely related. These results showed that allelic variation had been reduced considerably through the half-sib RS program, and very little reduction occurred after that in the S_2 RS program. Even so, approximately 30% of the loci showed changes in allele frequencies that varied more than could be explained by genetic drift alone. Particularly, *Ill. Hy*—the only progenitor in common between BSSS and BSCB1—had several alleles that deviated substantially from neutrality (Guimarães, 2001).

Hagdorn et al. (2003) examined data from 105 RFLP loci in 16 BSSS progenitors, 12 BSCB1 progenitors, 18 inbred lines derived from BSSS, and 7 inbreds derived from BSCB1. The objective was to determine the contributions of progenitors to derived lines and evaluate differences in genetic diversity among lines derived from early and advanced selection cycles. In keeping with the results mentioned above, the progenitors were genetically very broadly based and highly variable at individual loci, with 27% unique alleles in the

progenitor groups. But overall, the BSSS progenitors were not highly divergent from the BSCB1 progenitors.

Within the BSCB1 group, no single progenitor was identified as a contributor of a majority of it's alleles to the derived lines (Hagdorn et al., 2003). In BSSS, though, progenitor *CI540* contributed more alleles to the derived lines than any other BSSS progenitor. All progenitors appeared to have made some contribution to the derived lines, because each one had a few alleles that appeared in the derived lines, but with a significant increase in frequency. But because not all of the alleles are unique to a single progenitor, it is impossible to determine the exact origin of each descendant allele (Hagdorn et al., 2003).

Another key finding of Hagdorn et al. (2003) related to the range of alleles recovered in the derived lines. The authors reported that 75% of the alleles in BSSS progenitors and 67% of alleles in BSCB1 progenitors were seen in their respective derived line groups. Therefore the derived lines have captured a majority of the genetic variation that was present in the progenitors.

Statistical Methods

We are interested in changes in genetic structure between inbred progenitors of populations and inbred lines derived from populations, not the populations per se. Therefore, we genotyped the inbred progenitors of the populations and inbred lines derived from various improved cycles of the populations. In the case of BSSS, it is likely that we have not identified all possible sources of the alleles for the C0 population due to missing two of the progenitors (*CI167* and *F1B1*). The parents of *F1B1* are available, however, which is quite useful in this context. But uncertainty remains about what alleles may have been present in

CI167. Furthermore, three other progenitors—*I159*, *Ind. Tr 9-1-1-6*, and *A3G-3-1-3*—were inadvertently omitted from the laboratory procedures, so their SSR data are missing as well. For BSCB1 the situation is more promising since all twelve progenitors are available for genotyping. The point is that in a study of this nature we are not actually "sampling" the population, but rather measuring most (in BSSS) or all (in BSCB1) of the alleles that formed the original populations.

In a similar manner, the derived inbreds are not a random sample of the various cycles of selection, either. They represent the results of selection within the populations. Therefore the genotype of each derived line consists of alleles that were present in the cycle of origin, but give no indication of the abundance of those alleles in the population. Furthermore, an inbred line derived from any given cycle may not be finished and released until several years have passed. During the elapsed time the populations of origin will have continued to undergo evaluation and selection. This continuation will presumably, and perhaps substantially, alter the allele frequencies of the underlying population. So by genotyping inbreds we cannot expect to compare results with the current population cycles, but only with the cycles of origin—and even that is somewhat of a meaningless comparison because there are so few inbred lines derived from a given cycle of the population. By way of example, the reciprocal recurrent selection program between BSSS and BSCB1 is currently in cycle 16. The most recent derived lines included in this study are from C11 in BSCB1 and C9 in BSSS. (There are also two inbreds from C5 of BS13, which would correspond approximately with C12 of BSSS.)

We can, though, consider the derived lines as a group, and their genotypes to be information about which progenitor alleles may have some agronomic significance in the

populations. Inbreds derived from later cycles may contain different alleles than inbreds derived from earlier cycles. The difficulty lies in trying to discover the reason for these allelic changes. The absence of certain alleles in advanced cycle inbreds does not necessarily indicate the absence of those alleles in the population per se. With only one or two derived lines to give us the information about a specific cycle of the populations, our best hope is to try to locate alleles that seem to be present in high frequencies among the derived lines as a group. In this situation we might then conclude that these alleles are linked to alleles that have some positive agronomic impact, and that our evaluation methods are adequate to measure, detect, and select for that (those) trait(s).

Since the inbred lines evaluated in this study cannot be used to give us information about the number of alleles per locus in their cycle of origin in the populations per se, the only way to approximate that statistic is to consider the derived inbred lines collectively as a sample of the population. Conceptually, this would be a sample of either the BSSS or BSCB1 populations in general, not a sample of any particular cycle of the populations. But another aspect of this approach to the RRS program is that it is difficult to conclude that any genetic changes are due to a specific cause, such as selection or random genetic drift. Since the selection program is not replicated, any observations can be considered to be a single sample of all possible random genetic drift events in an RRS program.

Gene diversity, heterozygosity, and polymorphism information content (PIC)

Heterozygosity refers to the number of heterozygous individuals in a population. According to Weir (1996), heterozygosity is an important measure of variation in populations, but may not be the best measure for selfing species. In these cases gene diversity

is then a more appropriate measure. Since we examined inbred lines in this study, we will use the gene diversity statistic to measure not the presence of heterozygotes, but to measure the presence of different homozygotes. Since the two measurements are related, both will be discussed.

The heterozygosity is estimated at a single locus using the formula

$$\hat{H}_l = 1 - \sum_{u=1}^{k} \tilde{P}_{luu} \qquad \text{or the equivalent} \qquad \hat{H}_l = \sum_{u \neq v} n_{luv} / n$$

Where $\tilde{P}_{luu}$ is the estimated frequency of homozygotes of any allele u at locus l, and n_{luv} is the number of heterozygotes, $u \neq v$, at locus l, in a sample of size n. Over m multiple loci, the heterozygosity measure becomes an average over loci as

$$\hat{H} = \frac{1}{m} \sum_{l=1}^{m} \hat{H}_l$$

Gene diversity, sometimes referred to as expected heterozygosity, is a related measurement from the sum of squares of allele frequencies (Weir, 1996). For random mated populations this measure will be close to the heterozygosity, but gene diversity is more appropriate for inbred populations, particularly when there are several different homozygous genotypes. If p_{lu} represents the frequency of an allele u at locus l, then

$$D = 1 - \sum_{u} p_{lu}^2$$

and averaged over m loci,

$$D = 1 - \frac{1}{m} \sum_{l} \sum_{u} p_{lu}^2$$

The polymorphism information content (PIC) is a measurement closely related to gene diversity. This measurement is defined as the probability that the marker genotype of a given progeny will allow deduction of which of the two marker alleles of the parent it received, assuming no crossing-over occurred (Guo and Elston, 1999). It is calculated as

$$PIC = 1 - \sum_{u} p_u^2 - \sum_{u=1}^{n-1} \sum_{v=u+1}^{n} 2 p_u^2 p_v^2$$

where p_u is the population frequency of allele u, and there are n alleles at a locus.

Polymorphism information content is used more often in linkage studies, and is probably not useful for us in this study. In future studies this statistic may be more useful in trying to track short haplotypic segments from progenitors into various cycles of the populations, and possibly into derived lines. In our study a major difficulty in trying to follow alleles from progenitor to progeny is that so few of the marker alleles that were unique to a single progenitor have been found in the derived lines. Therefore most of the alleles have multiple possible origins. Only 10.5% of the alleles in SSL, and 14.2% in CBL, are unique to a single progenitor. The PIC may be useful in detecting nonunique alleles that are linked to unique alleles, thus identifying a possible haplotypic segment. This aspect of the research will be left to future students.

Hardy-Weinberg equilibrium (HWE)

The principle of HWE is used to relate the genes of parents to the genotypes of progeny (Falconer and Mackay, 1996). The Hardy-Weinberg law states that, in large populations under conditions of random mating (no selection, no mutation, and no migration), the allele frequencies and the genotype frequencies are constant from one

generation to the next. If *A1* and *A2* represent both alleles at a biallelic locus, and p and q represent the frequencies of the alleles, then $p+q=1$. Under the random-mating conditions stated above, the *A1A1*, *A1A2*, and *A2A2* genotype frequencies are related to the allele frequencies by the following: $f(A1A1)=p^2$, $f(A1A2)=2pq$, $f(A2A2)=q^2$, and $p^2+2pq+q^2=1$.

The implications of this relationship are clear: If genotype frequencies differ from this relationship, then one or more of the assumptions of random mating are being violated. Therefore in molecular studies where both allele and genotype frequencies are known, the data can be examined for deviations from HWE. This measurement cannot be utilized in the current study, again because of the nature of our subject. Our derived inbred lines are not samples of a random mating population. They represent selected lines derived from populations that have undergone selection themselves, and therefore we cannot reasonably assume that any of the genotypes or alleles found among the derived lines exhibit HWE.

Simulations

One of the main objectives of our study was to determine if we could detect significant changes in allele frequencies and gene diversity from the progenitor groups to the derived line groups. We determined that changes in allele frequencies would probably be very difficult to interpret. Simulations of random drift effects on allele frequencies probably will not provide the answers. Hinze (2003) provided graphs to show the possible outcomes of 1000 simulations of the RRS program, including 95% confidence intervals (CI) on the results. But population genetic theory already tells us that the probability of any allele reaching fixation is equal to the allele frequency (Hedrick, 2000; Falconer and Mackay,

1996). Likewise, the probability of an allele extinction is equal to the inverse of the fixation probability. In the Hinze (2003) drift simulations for an allele with a beginning frequency of 0.5, the 95% CIs include 0 and 1 by C7 of the RRS program. For alleles with an initial frequency of 0.1, the 95% CI includes 0 by C1, and is over 0.5 by C7. Essentially what this tells us is that the majority of observed allele frequency changes may be due to random drift. Another obstacle preventing us from using this type of simulation is the nature of our data—we have inbred lines derived from various cycles of the populations, not samples of the populations per se. Therefore we cannot make specific conclusions about allele frequencies that are based on inbreds originating from different cycles of the populations.

Perhaps the best way to determine if any observed genetic changes might be something other than drift would be to examine the gene diversity among the derived lines rather than the allele frequencies. Based on the nature of our data, the best way for us to determine significance was to use simulations of our data set. The jackknife method could be used to provide estimates of the variance of our genetic parameter estimates, but gives little information about the distribution of those estimates (Weir, 1996). The bootstrap, however, can give a good approximation of the distribution of a parameter, and is not limited by the number of individuals or loci (Weir, 1996).

To perform the simulations, we attempted to mimic the RRS program for each marker, using as the starting point the number of alleles and their frequencies from our measurements of each progenitor group. We also simulated deriving an inbred from the various cycles that corresponded to the origins of the derived lines in our study. Gene diversity was then calculated among the group of derived lines for each marker locus for each run of the simulation. After 10,000 simulation runs for each marker, a probability

distribution was calculated for the gene diversity measurements from the simulated observations. The observed gene diversity was compared to the simulated diversity distribution. All observed diversity measurements that fell in or below the lower 10% tail of the simulated distribution were noted for follow-up study.

We used SAS (2003) to simulate the RRS program for each marker locus. The idea was to simulate what might be expected under conditions of random genetic drift. The methods of the real RRS that are pertinent to the simulation program are the number of individuals selected each cycle, and the method of recombination. Ten individuals were selected in C0 through C7, and 20 individuals were selected in each cycle thereafter. S_1 recombination was used from C0 through C5, and S_2 recombination has been used from C6 to the present.

To describe the simulation program, assume we are simulating a locus that contains four alleles in equal frequencies among the 12 members of the progenitors of BSCB1. Also assume that we have eight inbred lines derived from BSCB1 cycles, each one derived from a separate cycle from C0 to C7. The four observed progenitor alleles make up our hypothetical C0 population. During the first six cycles of RRS, with S_1 recombination and ten selections used to advance cycles, the simulation program would choose 20 alleles, with replacement, from the allele "pool". In this example, each of the four alleles from the progenitors was chosen with a probability of 0.25 (their beginning allele frequency in the hypothetical C0). Also, an allele was chosen to represent the inbred line derived from the C0. The probability for any allele to be chosen was equal to its frequency in the population.

During the simulation, the C(n+1) allele frequency is determined by the random sample of alleles chosen in C(n). Therefore, each run creates a new mix of alleles from the

previous generation, generating a new set of allele frequencies, which produces a new probability of choosing any particular allele to represent the advanced cycle. In our example, the process just described would occur for the first five cycles.

Beginning in cycle six, S_2 recombination was used in the RRS program. In this situation there is a probability of 0.5 that one of the individuals used in recombination is homozygous for an allele at the locus of interest. The simulation program dealt with this by performing a slightly different selection method. First, a random digit was chosen from a binary set. A "1" simulated that the S_2 individual chosen for recombination was homozygous for its allele at the locus of interest, and the chosen allele was then counted twice toward the total of 20 alleles among the selected individuals. A "0" meant the individual was heterozygous at the locus, and two alleles were chosen in a manner identical to that described above. As in previous cycles, an allele was chosen to represent the inbred line derived from the cycle, the probability of which was equal to the allele frequency.

In an attempt to reduce the effects of genetic drift, the number of individuals selected for cycle advancement was increased to 20 beginning in cycle eight. The program has been operated in this manner since that time: 20 selected individuals using the S_2 as the recombination generation. The simulation program operated as in cycles six and seven, but now chose 40 alleles instead of 20. (Note: In this example no alleles were needed to represent inbreds beyond C7, but many of our observed derived lines originated from later cycles.)

Gene diversity was calculated for the derived lines based on the alleles chosen to represent the inbred lines from the various cycles. This was done for each simulation run, yielding 10,000 gene diversity measurements, and their corresponding probability, for each marker. A cumulative probability was calculated for each diversity measurement. The

cumulative probability represented the probability of observing a gene diversity less than or equal to the diversity associated with a given cumulative probability. The gene diversity values calculated from our observed derived lines were compared with the cumulative probabilities of the simulated diversity measurements for each marker. If the cumulative probability of the observed gene diversity was less than 10%, the markers were chosen for further study.

A significant reduction in gene diversity means that there are fewer alleles at a locus than expected, considering the expectation is based on the RRS, the number of alleles and their frequencies among the progenitor lines, and random genetic drift. This reduction in allele diversity is a potential indicator of some effect on allele frequencies other than random genetic drift, possibly selection. If this happens to be true, then the marker locus may be linked to a favorable single-gene trait or quantitative trait locus (QTL).

Genetic distance

Differences and similarities in patterns of genetic variation can be the result of many factors. According to Hedrick (2000), two populations may be genetically similar because:

1) they were only recently separated into two populations;

2) gene flow occurred between them;

3) they were large populations, so there was little genetic drift;

4) similar selection pressures affected loci similarly in both populations.

If two populations are genetically quite different, the possible reasons are essentially the opposite of those just given for similarity. More than one of these factors, and possibly all, may be important, depending on the populations under study.

Many genetic distance measurements have been proposed, often with only minor differences between them when differences between populations are small (Hedrick, 2000). When differences become large, though, there may be substantial differences between different genetic distance measurements on the same data set.

The most commonly used distance measurement is that of Nei (1972). In cases of no differential selection (neutrality), and when all new mutations result in new alleles, this distance will increase linearly with time Hedrick (2000). For our study, a more appropriate distance measurement is Nei's 1978 method, which is designed to overcome a bias in the measurement when sample sizes are small. The 1972 distance calculation will be described first, followed by the adjustment to account for small sample size. All formulas and variable descriptions are taken or modified from Nei (1972; 1978).

Let X and Y represent two diploid populations in which multiple alleles segregate at a locus. The frequencies of the i^{th} alleles in X and Y are given by x_i and y_i, respectively. Then the probability of identity of two randomly chosen alleles is $j_X = \sum x_i^2$ in population X, and $j_Y = \sum y_i^2$ in population Y. The probability of identity of two alleles, one from X and one from Y, is given by $j_{XY} = \sum x_i y_i$. The normalized genetic identity of alleles between X and Y at a single locus is $I_J = j_{XY} / \sqrt{j_X j_Y}$. The normalized identity of alleles between X and Y over all loci is $I = J_{XY} / \sqrt{J_X J_Y}$, where J_X, J_Y, and J_{XY} are the arithmetic means of the corresponding single-locus values. The genetic distance between X and Y is then defined as $D = -\ln\left[J_{XY} / \sqrt{J_X J_Y} \right] = -\ln I$.

The sample gene identities of J_X, J_Y, and J_{XY} are understood to be estimates of the population gene identities of G_X, G_Y, and G_{XY}. Later, Nei (1978) discovered that in the case of a small sample size (small *n*), the genetic distance measurement as described above will have a tendency to overestimate the true distance. When $D = 0, G_X = G_Y = G,$ and $n_X = n_Y = n$, then the expectation of $\hat{D}$ is approximately $(1-G)/(2nG)$. The result is that even when two populations are genetically identical, the sample genetic distance can be larger than 0 when the sample size is small.

To overcome this problem an unbiased estimate of *D* may be found by using the unbiased estimates of G_X and G_Y for J_X and J_Y. The distance formula then becomes $\hat{D} = -\ln\left[\hat{G}_{XY} / \sqrt{\hat{G}_X \hat{G}_Y}\right]$, where $\hat{G}_X$ and $\hat{G}_Y$ are the multilocus averages of $(2n_x J_X - 1)/(2n_x - 1)$ and $(2n_Y J_Y - 1)/(2n_Y - 1)$, and $\hat{G}_{XY} = J_{XY}$.

Nei (1978) also notes that in cases of large sample sizes, on rare occasions $\hat{D}$ may be negative, though the absolute value should not be large. This is reported to be due to sampling error, and Nei suggests converting any negative values to 0.

Rogers (1972) proposed an alternative method of calculating genetic distance. The formula for Rogers's distance (RD) is $\mathrm{RD} = \frac{1}{m}\sum_{i=1}^{m}\sqrt{\left(\frac{1}{2}\right)\sum_{j=1}^{a_i}\left(p_{ij} - q_{ij}\right)^2}$ where

m = number of loci

a_i = number of alleles at the i^{th} locus

p_{ij} and q_{ij} = allele frequencies of the allele j at the i^{th} locus in a respective pair or group of lines.

For homozygous lines the RD measurement corresponds with the proportion of marker loci which differ between the two lines (Hagdorn, 2003).

Rogers (1991) compared the usefulness of various distance measurements when developing phylogenetic trees from allele frequencies. At issue was the way different methods treated heterozygosity in populations, or how sample sizes might affect measurements. The maximum single-locus distance between any two populations is 1, and this should occur if the populations share no alleles. Rogers's distance, however, will assign a distance of 1 only when the two populations are *fixed* for different alleles. The Rogers's distance will be less than 1, and the magnitude of the maximum distance decreases as heterozygosity increases (Rogers, 1991). This has been considered to be a flaw in the Rogers's distance method, leading to a modified Rogers's distance from Wright (1978), among others.

Sample sizes and sampling error also play a critical role in estimating genetic distances. Due to sampling error, even samples drawn from the same population will rarely be identical, and in such cases will overestimate the true genetic distance of zero. Rogers's distance is more sensitive to sample heterozygosity in populations, and tends to overestimate the true distance of zero by a much smaller margin than some other methods (Rogers, 1991).

When comparing Rogers's distance to Wright's modified Rogers's distance, Rogers states that under the modified form, the optimization equation for a given allele frequency is a function of the frequencies of other alleles at that locus *and* the alleles of all other loci as well. For the original Rogers's distance, the optimization equation for a given allele is dependent only on the frequencies of alleles at the same locus. Rogers indicates his preference for this method because the frequencies of alleles at a locus are mathematically

dependent, but those for different loci are independent. Even in the cases of alleles at linked loci, the nature of the interaction is unlikely to conform to the mathematical structure of the modified Rogers's distance or any other distance that does not assume independent loci. For our study, we will use the original Rogers's distance (1972).

Principal component analysis (PCA)

Principal component analysis is a method of reducing the dimensionality of multivariate data. In practice, that means PCA is used to describe the variation of multivariate data using a set of uncorrelated variables, each of which is a linear combination of the original variables (Everitt and Dunn, 2001). The new variables are derived in order of decreasing importance. The first principal component (PC) must therefore describe as much variation as possible of the original data. The objective is to determine if a few variables can be used to account for much or most of the variation in the data. If so, the PCs can then be used to summarize the data with little loss of information.

The use of multivariate statistical methods requires that linear combinations of traits must be normally distributed. Therefore the diploid genotype data in our study was transformed following the methods described in Smouse and Williams (1982). Briefly, since the frequencies within a locus must sum to unity, there are $n-1$ independent pieces of useful information, where n represents the number of alleles at a locus. For each allele in our study then, the data are described using a per-locus frequency within each individual. An example of the transformation method is shown in Table 1. Smouse et al. (1982) have shown that linear combinations of such allelic scores over ten or more loci are indeed normally distributed, following the expectations of quantitative genetic theory (Kempthorne, 1969).

The following description of the definitions associated with, and methods used to calculate, PCs is summarized from Everitt and Dunn (2001) and Johnson and Wichern (1982). The first PC is the linear combination, y_1, of the original variables,

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p$$

whose sample variance is greatest for all coefficients $a_{11} \cdots a_{1p}$ (which is a vector, $\mathbf{a}_1$). A restriction is placed on the coefficients or else the variance of y_1 could be increased simply by increasing the elements of $\mathbf{a}_1$. One common restriction is that the sums of squares of the coefficients must equal unity, $\mathbf{a}_1'\mathbf{a}_1 = 1$.

The second PC is the linear combination, y_2, of the original variables,

$$y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \text{ (equal to } y_2 = \mathbf{a}_2'\mathbf{x})$$

which has the greatest variance now subject to two conditions: 1) $\mathbf{a}_2'\mathbf{a}_2 = 1$, for the same reason just described, and 2) $\mathbf{a}_2'\mathbf{a}_1 = 0$, so that y_1 and y_2 are uncorrelated. All succeeding PCs are determined in a similar manner, and subject to the constraints that $\mathbf{a}_i'\mathbf{a}_i = 1$, and $\mathbf{a}_j'\mathbf{a}_i = 0 \ (i < j)$.

The variance of y_1 is given by $Var(y_1) = Var(\mathbf{a}_1'\mathbf{x}) = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1$, where $\mathbf{S}$ is the variance-covariance matrix of the original variables. As mentioned, the objective in determining the first PC is to maximize this variance. A method commonly used for maximizing a function of several variables subject to constraints is called Lagrange multipliers. The end result is that $\mathbf{a}_1$ is the eigenvector of $\mathbf{S}$ corresponding to the largest eigenvalue. The Lagrange multiplier method is then used to get the second PC, again with the goal of maximizing the variance of

y_2. Again, the result is that $\mathbf{a}_2$ is the eigenvector of $\mathbf{S}$ corresponding to the second largest eigenvalue. This procedure is repeated, and always the j^{th} PC is defined by the eigenvector of the j^{th} largest eigenvalue.

The usefulness of any particular PC can be determined by the proportion of the total variance for which it accounted. The total variance of all PCs will equal the total variance of all variables of the original data. When all variables are uncorrelated, there will be as many PCs as there are variables to fully reproduce the total system variability. If traits are highly correlated with one another, the variation can be reduced to a very few PCs (Westfall and Conkle, 1992). Even in cases of uncorrelated variables, though, typically a majority of the variation can be accounted for by a relatively small number of components. The proportion of the total variation accounted for by the j^{th} PC is

$$t = \frac{\lambda_j}{trace(\mathbf{S})}$$

where λ_j is the variance of the j^{th} PC, and $trace(\mathbf{S})$ is the sum of the diagonal elements of the variance-covariance matrix $\mathbf{S}$.

In addition to using the variance-covariance matrix $\mathbf{S}$, PCs can also be derived from the corresponding values of the correlation matrix $\mathbf{R}$. In cases of multivariate data where the variables are of completely different types, the structure of the PCs derived from $\mathbf{S}$ will depend on the essentially arbitrary choice of units of measurement (Everitt and Dunn, 2001). If there are large differences in the variances of the variables, those variables with the largest variance will dominate the first few PCs. To overcome this problem, the PCs can be determined from $\mathbf{R}$, which is the equivalent of deriving them from the original variables after standardizing to have a unit variance. The correlations allow for differences in the

variances of the original variables, and avoid the problems of interpreting different measurement scales (Johnson and Wichern, 1982). The eigenvalues and eigenvectors of **R** do not generally correspond with those of **S**. According to Everitt and Dunn (2001), "choosing to analyse [sic] **R** rather than **S** involves a definite but possibly arbitrary decision to make the variables 'equally important'."

In our study we used the correlation matrix **R** to calculate the PCs for our SSR data. Using the covariance matrix causes markers with large variances to be more strongly associated with components with large eigenvalues and causes markers with small variances to be more strongly associated with components with small eigenvalues. Although the units in which our marker data are presented are identical for all markers (variables), we did not standardize the variables in any way to equalize variances. For comparisons sake, we also did PCA using the covariance matrix and compared the graphical results with the previous output. The dispersion patterns and data separation were very similar, but with a slightly more compressed arrangement when the covariance matrix was used. Only the results from the correlation matrix will be presented in this dissertation.

References

Everitt, B.S., and G. Dunn. 2001. Applied multivariate data analysis, 2nd Ed. Arnold Pub., London.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. Longman Group Ltd., Essex, England.

Guimarães, P. 2001. Effects of a long-term recurrent selection program on the genetic structure of the BSSS maize population. PhD dissertation. Iowa State University, Ames, IA.

Guo, X., and R.C. Elston. 1999. Linkage information content of polymorphic genetic markers. Hum. Hered. 49:112-118.

Hagdorn, S., K.R. Lamkey, M. Frisch, P.E.O. Guimaraes, and A.E. Melchinger. 2003. Molecular genetic diversity among progenitors and derived elite lines of BSSS and BSCB1 maize populations. Crop Sci. 43:474-482.

Hedrick, P.W. 2000. Genetics of populations, 2nd Ed. Jones and Bartlett Publishers, Sudbury, MA.

Hinze, L.L. 2003. Molecular marker analysis of population genetic structure and progress from reciprocal recurrent selection in two Iowa maize (*Zea mays* L.) populations. PhD dissertation. Iowa State University, Ames, IA.

Johnson, R.A., and D.W. Wichern. 1982. Applied multivariate statistical analysis. Prentice-Hall, New Jersey.

Kempthorne, O. 1969. An introduction to genetic statistics. Iowa State University, Ames, IA.

Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 1997. Molecular genetic diversity after reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 37:416-423.

Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 1999. Temporal changes in allele frequencies in two reciprocally selected maize populations. Theor. Appl. Genet. 99:1166-1178.

Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 2000. Hardy-Weinberg and linkage equilibrium estimates in the BSSS and BSCB1 random mated populations. Maydica. 45:243-255.

Lamkey, K.R. Fifty years of recurrent selection in the Iowa Stiff Stalk Synthetic maize population. Maydica. 37:19-28.

Messmer, M.M., A.E. Melchinger, M. Lee, W.L. Woodman, E.A. Lee, and K.R. Lamkey. 1991. Genetic diversity among progenitors and elite lines from the Iowa Stiff Stalk Synthetic (BSSS) maize population: comparison of allozyme and RFLP data. Theor. Appl. Genet. 83:97-107.

Nei, M. 1972. Genetic distance between populations. Am. Nat. 106:283-291.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics. 89:583-590.

Neuhausen, S.L. 1989. A survey of Iowa Stiff Stalk Synthetic parents, derived inbreds and BSSS(HT)C5 using RFLP analysis. Maize Genet. Coop. Newsletter. 63:110-111.

Penny, L.H., and S.A. Eberhart. 1971. Twenty years of reciprocal recurrent selection with two synthetic varieties of maize (*Zea mays* L.). Crop Sci. 11:900-903.

Rogers, J.S. 1972. Measures of genetic similarity and genetic distance. Studies in genetics. VII. Univ. Texas Publ. 7213:145-153.

Rogers, J.S. 1991. A comparison of the suitability of the Rogers, modified Rogers, Manhattan, and Cvalli-Sforza and Edwards distances for inferring phylogenetic trees from allele frequencies. Syst. Zool. 40:63-73.

Rouse, J.R., L.L. Hinze, and K.R. Lamkey. 2003. Selection history and developments from the BSSS and BSCB1 maize populations. *In* International Symposium on Plant Breeding poster presentations [CD-ROM]. CIMMYT, Mexico City, Mexico.

Russell, W.A., L.H. Penny, A.R. Hallauer, S.A. Eberhart, G.E. Scott, W.D. Guthrie, and F.F. Dicke. 1971. Registration of maize germplasm synthetics. Crop Sci. 11:140-141.

SAS Institute. 2003. SAS procedure guide computer program, version 9.1. SAS Institute, Cary, NC.

Smouse, P.E., and R.C. Williams. 1982. Multivariate analysis of HLA-disease associations. Biometrics. 38:757-768.

Smouse, P.E., R.S. Spielman, and M.H. Park. 1982. Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. Am. Nat. 119:445-463.

Sprague, G.F. 1946. Early testing of inbred lines of corn. J. Am. Soc. Agron. 38:108-117.

Troyer, A.F. 2004. Background of U.S. hybrid corn II: Breeding, climate, and food. Crop Sci. 44:370-380.

Waples, R.S. 1989. Temporal variation in allele frequencies: testing the right hypothesis. Evolution. 43:1236-1251.

Weir, B.S. 1996. Genetic data analysis II: Methods for discrete population genetic data. Sinauer Assoc., Sunderland, MA.

Westfall, R.D., and M.T. Conkle. 1992. Allozyme markers in breeding zone designations. New Forests. 6:279-309.

Wright, S. 1978. Evolution and the genetics of populations, Volume 4. Variability within and among natural populations. University of Chicago Press, Chicago, IL.

Zuber, M.S., and L.L. Darrah. 1981. 1979 U.S. corn germplasm base. Proc. 35th Ann. Corn & Sorghum Res. Conf. ASTA. pp. 234-249.

Table 1. Data transformation example for an n = four-allele, diploid locus. The individual genotypes listed are transformed into a frequency vector **Y** consisting of $(n-1)$ members.

Genetic variable in output vector	Genotype of individual									
	11	12	13	14	22	23	24	33	34	44
Y_1	1	½	½	½	0	0	0	0	0	0
Y_2	0	½	0	0	1	½	½	0	0	0
Y_3	0	0	½	0	0	½	0	1	½	0

CHAPTER 3: DETECTING SELECTION IN THE MAIZE GENOME BY USING GENETIC DIVERSITY

A paper to be submitted for publication in *Crop Science*
Jim R. Rouse, Lori L. Hinze, Jode W. Edwards, and Kendall R. Lamkey

Abstract

Forty-six inbreds related to Iowa Stiff Stalk Synthetic (BSSS) and Iowa Corn Borer Synthetic #1 (BSCB1) were assayed for polymorphism at 227 microsatellite loci. The inbreds consisted of progenitors of BSSS and BSCB1 as well as elite lines derived from those populations. Diversity statistics were used to estimate genetic variability among the derived lines, and to locate regions of the maize genome that have changed as a result of artificial selection. The four groups of germplasm were labeled CBP and SSP for the progenitors of BSCB1 and BSSS, respectively, and CBL and SSL for the lines derived from BSCB1 and BSSS, respectively.

With 227 simple sequence repeat (SSR) markers, there was an average of 22.7 markers per chromosome (range of 15-28). There were means of 3.5 and 3.4 alleles per locus among the CBP and SSP, respectively, and 2.3 and 2.5 alleles per locus among the CBL and SSL. Twenty-four percent of the SSRs were dinucleotide repeats, and 52% were trinucleotide repeats. Of the remaining SSRs, 13 had an unknown repeat length. The longest known SSR in this study was a septanucleotide repeat.

As expected, many more alleles were found in the progenitor groups than in the groups of derived lines. CBL showed only 60% of the alleles found in CBP, while SSL had 66% of the alleles found in SSP. Supporting previous studies in this area, we found that 25.8% of the alleles in SSP were unique to a single inbred. In CBP, 31.8% of the alleles were

unique, a figure slightly higher than previous results in BSCB1 or BSSS. Of the unique alleles in both progenitor groups, 73% were not found in any of the derived lines.

Regions of the genome undergoing selection should show changes in allele frequencies that are beyond that expected by random genetic drift alone. There were 33 marker loci in BSSS and 18 marker loci in BSCB1 that exhibited reductions in gene diversity that can be attributed to artificial selection (P = 0.1).

Introduction

Iowa Stiff Stalk Synthetic (BSSS) and Iowa Corn Borer Synthetic #1 (BSCB1) maize populations have been in a reciprocal recurrent selection (RRS) program since 1949. These two populations have also been used as sources of inbred lines throughout the course of the RRS. Hallauer (1984) stated that BSSS is one of the most important source populations for lines with above average stalk quality and combining ability. Over the last 50 years, approximately 50 inbreds and 30 populations containing germplasm derived in whole or in part from these two populations have been released to the public (Rouse et al., 2003). Approximately two-thirds of the lines and populations are related to BSSS, with the remainder related to either BSCB1 or both source populations. Zuber and Darrah (1981) estimated that about 19% of the total hybrid seed needed to plant the 1980 U.S. maize acreage was made from inbred lines derived from BSSS. This was considered a minimum estimate, however, because it did not include related inbred lines or proprietary lines derived from BSSS.

The longevity of the RRS program between BSSS and BSCB1 and the popularity of the germplasm developed from the program have led to numerous studies involving these

two populations. Details of the RRS program and its effects on various genetic and phenotypic parameters of the two populations can be found in Penny and Eberhart (1971); Eberhart et al. (1973); Martin and Hallauer (1980); Smith (1983); Lamkey et al. (1991); Keeratinijakal and Lamkey (1993); Schnicker and Lamkey (1993); and Holthaus and Lamkey (1995). There have also been numerous studies to examine the genetic structure, diversity, and other genetic parameters of these populations. For a more thorough review, see Messmer et al. (1991); Labate et al. (1997; 1999; 2000); Hagdorn et al. (2003); Guimarães (2001); and Hinze (2003).

The study described in this dissertation is related to some of the molecular studies just mentioned, and those described in the dissertation of Hinze (2003), and in Hinze et al. (in review). Hinze's studies involved 85 SSR markers on the progenitor lines of BSSS and BSCB1, the cycle 0 population (C0), and six advanced cycles (C1, C3, C6, C9, C12, and C15) of both populations. Hinze's main objective was to determine how the genetic structure of the populations has changed over time. The study described here is of a similar nature, but examines the genetic diversity of inbred lines derived from the populations, rather than the populations per se. Briefly, there are 227 SSR markers on the progenitors of BSSS and BSCB1, and on several inbred lines derived from these populations and the RRS program.

The goal of the RRS program is to improve both populations simultaneously, and maintain genetic variability within each population (Comstock et al., 1949). We examined measures of genetic diversity among and between the progenitors and the derived lines of BSSS and BSCB1 to estimate the genetic variability in these germplasm groups. Gene diversity was used to determine marker loci that have undergone changes due to artificial selection. A key element is that we are using marker information from inbred lines, not the

populations per se. Even though our lines are derived from the BSSS and BSCB1 populations, we are not trying to make claims about the genetic makeup of the populations.

Materials and Methods

Germplasm

There were 46 maize inbred lines, all from the U.S. Cornbelt, used in this study. They can be classified into the following four germplasm groups:

1) The 12 progenitors of the BSCB1 population, referred to as CBP;
2) Thirteen of the 16 total progenitors of BSSS, hereafter referred to as SSP;[†]
3) Eight inbreds derived from BSCB1, referred to as CBL;
4) Fourteen inbreds derived from BSSS, referred to as SSL.

While it seems like 47 inbreds were used, *Illinois Hy* was a progenitor of both populations. For this study, *Hy* was assayed once at each locus, and the results were duplicated for use in each progenitor group. Furthermore, four of the BSSS lines were derived from a recurrent selection program other than the RRS program described above. Hagdorn et al. (2003) deemed these lines to be related closely enough to be included as BSSS-derived lines. Statistical power to detect certain kinds of genetic change will be limited due to the small sample sizes of the derived line groups. A full list of the inbreds used, and their associated germplasm group, is given in Table 1.

[†] Of the original 16 BSSS progenitors, two (*CI617* and *F1B1*) have been lost. However the parents of *F1B1* (*Fe* and *IndB2*) are available, and when included with the other 14 BSSS progenitors thus constitute the 16 BSSS progenitors originally targeted for this study. Only 13 BSSS progenitors are included in this data set, though. Three lines, *I159*, *Ind. Tr 9-1-1-6*, and *A3G-3-1-3*, were inadvertently omitted from the laboratory procedures so no marker data are available on them.

Microsatellite genotyping

All molecular data were collected at Cornell University in the Institute for Genomic Diversity. For details see Hinze (2003) or Hinze et al. (in review). Briefly, a 5 cm leaf section from ~2-week old plants was collected and freeze dried until DNA extraction. A miniprep protocol using CTAB extraction buffer (Mitchell et al., 1997) was then used to extract genomic DNA from the leaf samples, and the SSR regions were amplified using PCR and fluorescent-labeled primers. PCR products were separated by size on a fragment analyzer (Applied Biosystems, Foster City, CA) equipped with GENESCAN 3.1 software. GENOTYPER 2.1 software (Applied Biosystems) was used to identify markers and their specific alleles. If no amplification products were seen in the gel image, PCR was rerun for that entry/SSR combination.

Two hundred eighty-four SSR markers were chosen for analysis based on their distribution throughout the maize genome. Marker locations (bins) are based on information from the Maize GDB (Lawrence et al., 2004). A bin is a relative genetic map position that represents about 20 centiMorgans on the genetic map of maize (Hinze, 2003). Fifty-seven markers were discarded from the data set due to low PCR amplification or ambiguous results, leaving 227 markers to be analyzed for this study.

Genetic measurements

We calculated allele frequencies, the number of alleles per locus, and gene diversity for each locus within each germplasm group. We also subdivided the marker information by repeat length to compare diversity measurements from dinucleotide repeat, trinucleotide

repeat, and greater length markers. These measurements were calculated using the software PowerMarker (Liu and Muse, 2003).

Gene diversity, sometimes referred to as expected heterozygosity, is calculated from the sum of squares of allele frequencies (Weir, 1996). For random mated populations this measure will be close to the observed heterozygosity, but gene diversity is more appropriate for inbred populations, particularly if there are several different homozygous genotypes. If p_{lu} represents the frequency of an allele u at locus l, then

$$D = 1 - \sum_{u} p_{lu}^2 ,$$

and averaged over m loci,

$$D = 1 - \frac{1}{m} \sum_{l} \sum_{u} p_{lu}^2 .$$

Simulations

One of the main objectives of this study was to determine if we could detect significant changes in gene diversity from the progenitor groups to the derived line groups. Reductions in gene diversity may highlight regions of the genome that have changed due to artificial selection pressure. Based on the nature of our data, the best way for us to determine significance was to use Monte Carlo simulations of our data set. To do this, we simulated the RRS program for each marker using the number of alleles and their frequencies from our measurements of each progenitor group. We also simulated deriving an inbred from the various cycles that corresponded to the origins of the derived lines in our study. Gene diversity was then calculated for each marker locus among the group of derived lines for each run of the simulation. After 10,000 simulation runs for each marker, a probability

distribution was calculated for the gene diversity simulations. The observed gene diversity was compared to the simulated diversity distribution. All observed diversity measurements that were in the lower 10% tail of the simulated distribution were noted for follow-up study.

We used SAS to simulate the RRS program for each marker locus. The goal was to simulate the possible allelic changes that could occur due to random genetic drift through the various cycles of selection. The beginning allele numbers and allele frequencies for each marker are provided by the observations from the CBP and SSP groups. Methods used in the real RRS program that we simulated are:

1) the number of individuals chosen to advance the population cycle (10 selections in C0 through C7, 20 selections from C8 to the present); and
2) the recombination generation (S_1 in C0 through C5, S_2 from C6 to the present).

Details of the S_1 and S_2 recombination methods in the actual RRS program can be found in Penny and Eberhart (1971), Lamkey et al. (1991), and Keeratinijakal and Lamkey (1993).

During the first six cycles of RRS (from C0 through C5), with S_1 recombination and 10 selections used to advance cycles, the simulation program would choose 20 alleles, with replacement, from the allele "pool". The allele frequency of any cycle C(n+1) was equal to the allele frequencies among the 20 chosen alleles from the C(n). For the simulation, the allele frequency in the progenitor group was assumed to be equal to the allele frequency in the C0 population. The probability of an allele being chosen from any given cycle of the population is based on the allele frequency in the population.

In addition to making selections for cycle advancement, the program also selected an allele to represent an inbred derived from the cycle. Inbred representation was done only for

those cycles that corresponded with the origin of a derived line in our data set. For example, two of the CBL originated from C0 of BSCB1, so our simulation program selected two alleles from the C0 to represent the alleles found in two (eventually) homozygous inbreds derived from the hypothetical C0 population.

S_2 recombination was used in the RRS program beginning in cycle six. With this method, a single S_1 plant is selfed, and the resulting S_2 line is recombined with other S_2 lines using a diallel or random mating block. We simulated the outcome of S_2 recombination using a two-step process. First, a binary code was used to determine whether a given S_1 plant contained alleles that were identical by descent (IBD, $f = 0.5$). Second, the simulation program chose two alleles from the population. If the binary step indicated alleles that were IBD, the chosen allele was counted twice toward the total of 20 selected alleles. If not, two alleles were chosen from the population in a manner identical to that described above. As in previous cycles, when appropriate, an allele was chosen to represent a derived inbred line from the cycle.

In cycle 8 the number of individuals selected for recombination to form the next cycle was increased to 20. The RRS program has been operated in this manner since that time: 20 selected individuals using the S_2 as the recombination generation. For our purposes, the simulation program operated as in cycles six and seven, but now chose 40 alleles instead of 20 to represent the 20 selected individuals.

Gene diversity was calculated for the simulated derived lines based on the alleles chosen to represent the inbred lines from the various cycles. This was done for each simulation run, yielding 10,000 gene diversity measurements, and their corresponding probabilities, for each marker. A cumulative probability was calculated for each diversity

measurement. The cumulative probability represented the probability of observing a gene diversity less than or equal to the diversity associated with a given cumulative probability. The gene diversity values calculated from our observed derived lines were compared with the simulated diversity distribution for each marker. Since we used a significance level of $P = 0.1$, if the cumulative probability of the observed gene diversity was less than 10% the markers were chosen for further study.

Using gene diversity to detect regions under selection raises questions about the risk of misclassification of some possibly significant events based on the nature of the simulation program. The concern regards not considering the potential temporal information contained in the alleles found in the derived lines. For example, lines derived from later cycles may all share the same allele, while lines derived from early cycles show one or two or three alleles that differ from the allele found in the later-derived lines. This may represent a significant genetic change that would not be detected when all derived lines are pooled to calculate gene diversity in the simulations.

To test this hypothesis, we partitioned our derived lines into subsets to include only lines derived from a "late" cycle of origin. In CBL, that meant we omitted data from two lines derived from C0 (*B42* and *B54*). The remaining six lines, derived from C7 through C11, composed the CBL Late Lines. In SSL, we omitted data from five lines from C0 (*B37*, *B67*, *B68*, *B69*, *B101*), one from C4 (*B64*), and one from C5 (*B73*) were dropped. The remaining seven lines, originating from cycles seven through twelve, composed the SSL Late Lines. As before, the CBL subset group remained separate from the SSL subset group.

Gene diversity in the Late Line subset groups was calculated for all marker loci. Loci showing a diversity of zero have the same allele across all inbreds within a group, and were

chosen for further examination. There were 44 and 24 markers that met this criterion in the CBL and SSL Late Line groups, respectively. The simulation program was used again, but only for these zero diversity markers. As before, the SSR data from the CBP and SSP lines were used as the beginning data for the simulations.

The simulation program was altered to select only six (for the CBL subset) or seven (for the SSL subset) inbreds from the various simulated cycles of selection that corresponded with the inbred cycles of origin. Gene diversity was then calculated for the selected inbreds for each run of the simulation. As before, a probability distribution was created, and the observed gene diversity measurement of zero was compared to the simulated probability distribution. Only one marker in CBL Late Lines (phi126) and one marker in SSL Late Lines (mmc0381) showed a zero gene diversity probability of less than 10%. Both of these markers had been identified as having a significant reduction in diversity based on the earlier simulations that included both early- and late-derived inbred lines in the sampling scheme.

As a further check on the simulation scheme, we compared the specific alleles that were found at each of the zero-diversity loci with their frequency in their respective progenitor groups. Of the 24 loci in SSL Late Lines with fixed alleles, all but two of them were the most frequent allele at that locus in the SSP. At the other two loci, the allele found in the late-derived lines was the second-most frequent of three alleles. In each case the SSP allele frequency was 0.31—not considered a rare allele. Furthermore, both of these marker loci were identified in the simulation program as having observed results that were considered significant at $P = 0.1$.

Results and Discussion

Summary statistics

Nearly 25% of the SSRs were dinucleotide repeats, while just over 50% were trinucleotide repeats (Fig. 1). Of the remainder, most were tetranucleotide repeats (13%). Thirteen of the SSRs (6%) had unknown repeat length when checked on the Maize Genetics and Genomics Database at http://www.maizegdb.org on 14 September, 2004. Chromosome five had only 15 SSR loci while chromosome two had the most with 28 loci (Fig. 2).

Over all inbreds genotyped we found 991 alleles at 227 loci, for an average of 4.4 alleles per locus (Table 2). As expected, the progenitor groups had more alleles per locus than the derived line groups. In CB, there was reduction of about 34%, from 3.5 alleles per locus in the CBP down to 2.3 alleles per locus in the CBL. In BSSS the reduction was about 26%, from 3.4 alleles per locus in SSP to 2.5 alleles per locus in SSL.

To search for differences based on repeat size, we partitioned the data set by the repeat length of the markers. Liu et al. (2003) described Type I and Type II SSRs as those consisting of dinucleotide repeats (Type I), and all repeats greater than dinucleotide (Type II). Since approximately 25% of our markers were dinucleotide repeats, and approximately 50% were trinucleotide repeats, we partitioned our data into three subsets: dinucleotide repeats, trinucleotide repeats, and all repeats longer than trinucleotides. The 13 markers with unknown repeat length were omitted from this part of the analysis.

In all groups, the dinucleotide repeats had the greatest number of alleles per locus, and the trinucleotide repeats had the fewest alleles per locus, though these numbers were quite similar to those of the markers with longer repeats.

Genetic diversity

Thirty-five of the 46 inbreds in this study had at least one heterozygous SSR locus. The 11 inbreds with no heterozygous loci were distributed among all four germplasm groups (data not shown). In the SSL group six of the 14 derived lines contained no heterozygous loci. Inbred *K230* in CBP and *Fe* in SSP were the progenitors with the most heterozygous loci, with 13 and 14 heterozygous loci, respectively. These results agree with those of Labate et al. (1997) and Hinze (2003), who also found that *K230* was highly heterozygous. Among the derived lines the most heterozygous loci were found in *B42* in CBL (10 loci) and *B89* in SSL (8 loci).

Considering the 227 SSR loci, 69 were found to be heterozygous in at least one of the inbred lines. The maximum number of heterozygous observations for any given marker was five. No marker was heterozygous in more than three lines within any germplasm group. Over all possible inbred-by-SSR locus combinations, there were 112 heterozygous observations, which amounts to less than 1.1% of the data.

Among the BSSS progenitors, 25.8% of the alleles are unique to a single inbred (Fig. 3). This compares favorably with the results of Messmer et al. (1991), Labate et al. (1997), and Hinze (2003), who all found approximately 25% unique alleles in the BSSS and BSCB1 progenitor groups. In the BSCB1 progenitors, though, nearly 32% of the alleles were unique (Fig. 3). This number is somewhat higher than the results just referenced, but similar to the nearly 31% unique alleles found by Guimarães (2001). Among the derived lines the percentage of unique alleles is lower than within the respective progenitor groups, at 16.6% in SSL and 28.7% in CBL (Fig. 4). This is not particularly surprising since the unique alleles in the progenitor groups have very low frequencies by their definition, many of these alleles

would be expected to be lost due to random genetic drift. Furthermore, if the alleles were linked to favorable genes that were under selection, we would expect to find them in more than just one of the derived lines.

Another important feature of these figures is what they reveal about allele frequencies—the majority of the alleles found in this study have very low frequencies. Sixty percent of the CBP and 58% of the SSP alleles have frequencies of 0.25 or less. Similarly, 57% of CBL and 58% of SSL alleles have frequencies lower than 0.37. In the progenitor groups only about 13% of the alleles had frequencies greater than or equal to 0.5, while about 30% of the alleles in the derived lines had frequencies in that range.

Table 3 shows the relationship between allele frequencies within CBP and their distribution among the CBL. The "Number of alleles" row contains the total number of alleles in each CBP allele frequency category. For example, there were 251 unique alleles among the CBP—alleles with a frequency of 0.083. The frequency corresponds to 1 inbred among the 12 progenitors carrying a given allele. The 251 unique alleles represent the 31.8% unique alleles mentioned previously. Likewise, 144 alleles had a frequency of 0.167, as they were found in only two of the 12 CBP lines. Eighty alleles were found in three of the 12 CBP lines, thus have a frequency of 0.25. Of all alleles detected in the CBP, only one was present in all 12 CBP lines (monomorphic among the CBP).

The columns of Table 3 reveal the distribution of alleles in CBL. Among all unique alleles from CBP, 183 (73%) were not found in any of the CBL, while 34 were found in only one of the CBL. The single allele that was present in all 12 of the CBP was also found in all eight CBL. Labate et al. (1997) defined novel alleles as alleles present in the populations per se that were not found in the progenitors. There were 37 alleles undetected in CBP that were

present in the CBL, with 26 (70%) being found in only one of the CBL (Table 2). Labate et al. (1997) stated that the novel alleles may represent contamination from an outside source. It is also possible that some of the novel alleles are the result of SSR scoring errors or SSR mutations.

Table 4 contains the corresponding data for the SSP and SSL groups. Again, the table shows that of the 197 unique alleles from SSP, 144 (73%) were not detected in the SSL. In both BSSS and BSCB1, in addition to the drastic loss of rare alleles, nearly half (47%) of the alleles that were found in only two progenitors were not found in any derived lines. There are more novel alleles in the SSL, but this is not surprising since we are missing marker information from three of the SSP. Therefore we expected to find alleles in the SSL that are not present in the SSP data set.

Alleles that were more abundant in the progenitors tend to be more abundant among the derived lines. Of course, that does not necessarily mean that the abundant alleles are more important than the less frequent alleles. The difficulty in interpreting the allelic distribution lies in finding changes that might be due to the forces of selection rather than random genetic drift.

Detecting regions under selection

A significant reduction in gene diversity means that allele frequencies are more skewed at a locus than expected, based on random genetic drift within the RRS, and the number of alleles and their frequencies among the progenitor lines. This reduction in allele diversity may indicate that a marker locus is influenced by some effect other than random

genetic drift. If this happens to be true, then the marker locus associated with that allele may be linked to a gene that has some influence on a trait of agronomic interest.

After comparing our observed genetic diversity in the derived lines with the simulated diversity distributions, we found 18 markers in BSCB1 and 33 markers in BSSS that had undergone changes significantly different from those expected due to random genetic drift (Table 5). The significant markers and related information are given in Table 6. Three of the markers were common between both BSCB1 and BSSS. At this point we are interested in two aspects of the marker information: 1) Are there regions of the genome that appear to be undergoing selection in both groups? 2) Do any of these regions correspond with the regions found by Hinze (2003) that may be undergoing selection?

The three common markers represented bins 2.08 (markers bnlg1233 and mmc0381) and 6.0 (marker umc1002). In bin 2.08, each derived line group contained an additional marker that showed a decrease in gene diversity. Thus, either these markers are very closely linked to each other, or this region contains one or more loci under selection. Other regions of the genome that had significant loci in both populations are bins 2.04, 3.07, 4.11, 5.04, and 9.04. The markers showing changes were not the same between the groups, but there was a least one significant marker in each of these bins in each population.

In the CBL Late lines there were 44 loci with a fixed allele. Twelve of those contained an allele other than the CBP most-frequent allele. Five of the alleles had starting frequencies of 0.33, two had starting frequencies of 0.25, three had frequencies of 0.18, and the other two alleles, with the lowest frequency, were unique alleles in the CBP. In the case of the unique alleles, neither locus had been previously identified in the simulation program as having undergone a significant change.

This result led to another aspect of the simulation analysis that we had not yet performed—calculating the probability of a particular allele being found in a given subset of derived lines. To do this, we used the same simulation program for the cycles of selection and derived lines from the later cycles to match the lines in our data set. To determine the probability of an allele with any particular frequency among the progenitors being found in each of the late-derived lines, we ran the simulations again, but with different starting allele frequencies. We made up a hypothetical progenitor data set containing six biallelic loci in CB, and eight biallelic loci in SS. For CB, since there are 12 progenitors, a unique allele has a frequency of 0.083 (1 out of 12 progenitors, or 2 out of 24 alleles). The next largest allele frequency (neglecting heterozygotes) is 2 out of 12, then 3 out of 12, and so on up to a frequency of 0.5. Likewise for SS, but with 16 progenitors we use eight hypothetical loci to represent frequencies from 1 out of 16 up to 8 out of 16 (0.0625 to 0.5).

Using the simulation program to "derive" an inbred from the cycles that match the origin of our derived lines, we then calculated how often a particular allele appeared in the derived lines. For BSCB1, there was a probability of 0.006 that a unique progenitor allele would be present in each of the six late-derived lines. While this seems very low, it must be considered in the context of this study—there were 251 unique alleles in the CBP. Based on these numbers, the expectation is that 1.5 unique alleles (251 * 0.006 = 1.5) would be present in all six of the CBL Late Lines. In our study, we found two unique alleles among the CBP that were present in all six of the CBL Late Lines.

An alternative method of examining these data is analogous to a comparison-wise versus experiment-wise error rate. If we define "success" as the probability of finding a unique allele in all six late-derived lines, then:

Prob(success) = 0.006

Prob(not success) = 1 – Prob(success) = 0.994

Prob(no success in 251 trials) = $(1 - \text{Prob(success)})^{251} = (0.994)^{251} = 0.221$

Prob(at least one success in 251 trials) = 1 - Prob(no success in 251 trials) = 0.779

These results indicate that, based on our simulations and the design of the RRS program, nearly 78% of the time we examine 251 unique alleles among the CBP, we would expect at least one unique allele to be fixed in all six of the late-derived lines.

Because of these results we concluded our simulation program was robust enough that we are not seriously compromising the quality of our reported results. Nevertheless, we acknowledge there is still a possibility that some loci may have undergone some changes due to selection and yet remain undetected. Known differences between our simulations and the real RRS program include: several generations of random mating occurred in each population before officially declaring the C0; recombination in many cycles included a diallel of the selections followed by a season of random mating; and recombination in some cycles consisted of only a diallel. The simulation did not model these factors because:

1) The number of random mated generations prior to C0 formation is unknown in BSSS.
2) The population sizes used in the random mating generations is unknown.
3) It is uncertain which recombination cycles used a diallel only, and which used a diallele with an additional generation of random mating.

The implication of these differences is that the simulation program is probably underestimating the effects of drift, but gives a preliminary analysis for this type of data examination.

Bin 4.11 was identified by Hinze (2003) in both populations as having deviations significantly different from random genetic drift. Other regions from Hinze (2003) that corresponded with those we found were bins 1.03, 2.04, 2.08, and 5.04 in BSSS, and 6.0 and 9.04 in BSCB1.

From the standpoint of associating agronomic traits with these markers, this information will help prioritize the regions and/or markers for further testing. Based on our data and that of Hinze (2003), the highest priority regions would seem to be 2.08 and 4.11—identified in both studies and in both populations. The next priority should go to those regions identified in both studies, and in both populations in at least one of the studies, such as regions 2.04, 5.04, 6.0, and 9.04. The third priority is regions identified by one marker in both populations, or by more than one marker in the same population. These are regions 1.03, 3.07, 3.09, 6.05, 9.03, and 10.04.

Labate et al. (1999) also observed regions of the genome that had undergone changes deemed to be significantly different from drift. Like our study, not all regions of the genome were equally represented by markers, but there were still some overlapping results. In particular, Labate et al. (1999) also identified region 2.04 in both populations, and 5.04 and 10.04 in BSCB1. Labate et al. (1999) lacked markers in many of the other regions in which we found significant loci.

The remaining regions identified in our study may also provide a reasonable starting place to search for genes of interest in the RRS program. The priorities mentioned above are stated simply as an interpretation of which regions might be most likely to yield successful outcomes in a phenotypic mapping study.

References

Eberhart, S.A., S. Debela, and A.R. Hallauer. 1973. Reciprocal recurrent selection in BSSS and BSCB1 maize populations and half-sib selection in BSSS. Crop Sci. 13:451-456.

Guimarães, P. 2001. Effects of a long-term recurrent selection program on the genetic structure of the BSSS maize population. PhD dissertation. Iowa State University, Ames, IA.

Hagdorn, S., K.R. Lamkey, M. Frisch, P.E.O. Guimaraes, and A.E. Melchinger. 2003. Molecular genetic diversity among progenitors and derived elite lines of BSSS and BSCB1 maize populations. Crop Sci. 43:474-482.

Hallauer, A.R. 1984. George F. Sprague: 60 years of corn research. Maydica. 29:351-355.

Hinze, L.L. 2003. Molecular marker analysis of population genetic structure and progress from reciprocal recurrent selection in two Iowa maize (*Zea mays* L.) populations. PhD dissertation. Iowa State University, Ames, IA.

Hinze, L.L., S. Kresovich, J. Nason, and K.R. Lamkey. (In review.) Population genetic diversity in a maize reciprocal recurrent selection program. Crop Sci.

Holthaus, J.F., and K.R. Lamkey. 1995. Response to selection and changes in genetic parameters for 13 plant and ear traits in two maize recurrent selection programs. Maydica. 40:357-370.

Keeratinijakal, V., and K.R. Lamkey. 1993. Responses to reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 33:73-77.

Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 1997. Molecular genetic diversity after reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 37:416-423.

Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 1999. Temporal changes in allele frequencies in two reciprocally selected maize populations. Theor. Appl. Genet. 99:1166-1178.

Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 2000. Hardy-Weinberg and linkage equilibrium estimates in the BSSS and BSCB1 random mated populations. Maydica. 45:243-255.

Lamkey, K.R., P.A. Peterson, and A.R. Hallauer. 1991. Frequency of the transposable element *Uq* in Iowa stiff stalk synthetic maize populations. Genet. Res. 114:2561-2568.

Lawrence, C.J., Q. Dong, M.L. Polacco, T.E. Seigfried, and V. Brendel. 2004. MaizeGDB, the community database for maize genetics and genomics. (Available at http://maizegdb.org). Nucleic Acids Res. 32:D393-D397.

Liu, K., and S. Muse. 2003. PowerMarker: New genetic data analysis software. Ver. 3.0. Free program distributed by the authors over the internet from http://www.powermarker.net.

Liu, K., M. Goodman, S. Muse, J.S. Smith, E. Buckler, and J. Doebley. 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics. 165:2117-2128.

Martin, J.M., and A.R. Hallauer. 1980. Seven cycles of reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 20:599-603.

Messmer, M.M., A.E. Melchinger, M. Lee, W.L. Woodman, E.A. Lee, and K.R. Lamkey. 1991. Genetic diversity among progenitors and elite lines from the Iowa Stiff Stalk

Synthetic (BSSS) maize population: comparison of allozyme and RFLP data. Theor. Appl. Genet. 83:97-107.

Mitchell, S.E., S. Kresovich, C.A. Jester, C.J. Hernandez, and A.K. Szewc-McFadden. 1997. Application of multiplex PCR and fluorescence-based, seme-automated allele sizing technology for genotyping plant genetic resources. Crop Sci. 37:617-624.

Penny, L.H., and S.A. Eberhart. 1971. Twenty years of reciprocal recurrent selection with two synthetic varieties of maize (*Zea mays* L.). Crop Sci. 11:900-903.

Rouse, J.R., L.L. Hinze, and K.R. Lamkey. 2003. Selection history and developments from the BSSS and BSCB1 maize populations. *In* International Symposium on Plant Breeding poster presentations [CD-ROM]. CIMMYT, Mexico City, Mexico.

Schnicker, B.J., and K.R. Lamkey. 1993. Interpopulation genetic variance after reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 33:90-95.

Smith, O.S. 1983. Evaluation of recurrent selection in BSSS, BSCB1, and BS13 maize populations. Crop Sci. 23:35-40.

Weir, B.S. 1996. Genetic data analysis II: Methods for discrete population genetic data. Sinauer Assoc., Sunderland, MA.

Zuber, M.S., and L.L. Darrah. 1981. 1979 U.S. corn germplasm base. Proc. 35th Ann. Corn & Sorghum Res. Conf. ASTA. pp. 234-249.

Table 1. Inbred lines analyzed in this study.

Line	Origin/Pedigree
	Progenitor lines of Iowa Corn Borer Synthetic #1 (CBP) ‡
A340	4-29 X 64
CC5	Golden Glow (renamed W23)
I205	Iodent
Ill. Hy	Illinois High Yield
K230	No origin listed, possibly from Midland Yellow Dent
L317	LSC
Oh07	CI.540 X Ill. L
Oh33	Clarage
Oh40B	Composite of 8 LSC lines
Oh51A	[(Oh56 X Wf9) Oh56]
P8	Palin Reid
R4	Funk Yellow Dent
	Progenitor lines of Iowa Stiff Stalk Synthetic (SSP) ‡
CI.187-2	Krug-Nebraska Reid strain X Iowa Gold Mine
CI.540	Illinois 2-ear Variety
I224	Iodent
Ill. 12E	Source unknown
Ill. Hy	Illinois High Yield
Ind. 461-3	Reid Medium (Duddleston No. 461)
Ind. AH83	Funks 176A
Ind. B2	Troyer Reid (Late Butler)
Ind. Fe	Troyer Reid (Early)
LE23	Illinois Low Ear
Oh3167B	Echelberger Clarage
Os420	Osterland Yellow Dent
WD456	Walden Dent

‡ All progenitor pedigrees are listed and referenced in Hagdorn et al. (2003).

Table 1. Continued.

Line	Origin/Pedigree
	Elite lines derived from Iowa Corn Borer Synthetic #1 (CBL)
B42	BSCB1C0
B54	BSCB1C0
B90	BSCB1(R)C7
B91	BSCB1(R)C8
B95	BSCB1(R)C7
B97	BSCB1(R)C9
B99	BSCB1(R)C10
B112	BSCB1(R)C11
	Elite lines derived from Iowa Stiff Stalk Synthetic (SSL)
B101	BSSSC0
B37	BSSSC0
B67	BSSSC0
B68	BSSSC0
B69	BSSSC0
B64	BSSS(R)C4
B89	BSSS(R)C7
B94	BSSS(R)C8
B105	BSSS(R)C9
B111	BSSS(R)C9
B73	BSSS(HT)C5
B84	BS13(S)C0
B104	BS13(S)C5
B110	BS13(S)C5
	BSSS progenitor lines that were omitted from this study
I159	Iodent
Ind Tr 9-1-1-6	Reid Early Dent (Troyer strain)
A3G-3-1-3	[BL345B X IAI129] (both strains of RYD)

Table 2. Allele information summarized by germplasm group.

Statistic	Group[†]				
	Overall	**SSP**	**SSL**	**CBP**	**CBL**
Sample size	47	13	14	12	8
Alleles	991	763	567	790	516
Alleles per locus	4.4	3.4	2.5	3.5	2. 3
Di-repeat SSR alleles/locus	7.1	4.6	2.9	4.8	2.6
Tri-repeat alleles/locus	3.6	2.9	2.3	3.0	2.2
> Tri-repeat alleles/locus*	3.8	3.3	2.5	3.3	2.2
Gene diversity	0.56	0.53	0.41	0.55	0.37
Di-repeat gene diversity	0.68	0.64	0.48	0.66	0.42
Tri-repeat gene diversity	0.51	0.50	0.37	0.52	0.37
> Tri-repeat gene diversity*	0.54	0.52	0.42	0.54	0.33
Group-specific alleles	205	71	25	88	21
Group-specific alleles/line	3.9	5.5	1.8	7.3	2.6
Group-specific alleles (%)	20.7	9.3	4.4	11.1	4.1
Line-specific alleles	171	64	20	73	14
Line-specific alleles (%)	17.3	8.4	3.5	9.2	2.5

[†] SSP = BSSS progenitors; CBP = BSCB1 progenitors; SSL = BSSS derived lines; CBL = BSCB1 derived lines.

* 13 loci had unidentified repeat length, were left out of these calculations.

Table 3. Alleles in BSCB1 progenitors and how they were distributed among the BSCB1 derived lines.*
Column headers denote the number of progenitor inbreds carrying identical alleles. Row headers denote the number of derived lines carrying those same alleles. Total number of alleles for each column are given in the bottom row. The first column of data shows there were alleles present in the derived lines that were not present in any of the progenitors. The first row of data indicates the number of alleles that were present in progenitors, but undetected in any of the derived lines.

		Number of BSCB1 progenitors (CBP) containing the same allele												
		0†	1	2	3	4	5	6	7	8	9	10	11	12
Number of CBL containing the same alleles	0	0	183	71	23	20	9	3	1	1	0	0	0	0
	1	26	34	30	22	14	10	6	3	3	0	0	0	0
	2	3	11	10	17	14	10	7	4	3	2	0	0	0
	3	4	10	11	5	10	6	9	5	2	0	2	0	0
	4	2	5	8	2	5	9	6	8	3	0	1	0	0
	5	1	3	8	5	5	8	10	4	2	3	2	0	0
	6	1	4	5	5	10	8	6	4	7	5	2	1	0
	7	0	1	1	1	6	3	5	4	3	4	2	0	0
	8	0	0	0	0	3	1	2	4	7	6	4	7	1
Number of alleles:		37	251	144	80	87	64	54	37	31	20	13	8	1

* This table shows there were 251 unique alleles (alleles found in a single progenitor only) in the CBP. Of those 251 alleles: 183 were not found in any CBL; 34 were found in 1 of the CBL; 11 were present in 2 CBL, etc. Similarly, 1 allele that was present in all 12 CBP was also found in all 8 CBL.

† This column represents novel alleles (Labate et al., 1997). There are 37 alleles that were not detected in any of the CBP, but were found in the CBL.

Table 4. Alleles in BSSS progenitors and how they were distributed among the BSSS derived lines.* Column headers denote the number of progenitor inbreds carrying identical alleles. Row headers denote the number of derived lines carrying those same alleles. Total number of alleles for each column are given in the bottom row. The first column of data shows there were alleles present in the derived lines that were not present in any of the progenitors. The first row of data indicates the number of alleles that were present in progenitors, but undetected in any of the derived lines.

		Number of BSSS progenitors (SSP) containing the same allele													
		0†	1	2	3	4	5	6	7	8	9	10	11	12	13
Number of SSL containing the same alleles	0	0	144	65	26	12	5	4	1	2	0	0	0	0	0
	1	31	18	16	13	3	10	3	0	0	0	0	0	0	0
	2	4	5	12	14	9	7	4	3	0	2	1	0	0	0
	3	8	11	9	10	5	8	8	6	0	1	2	0	0	0
	4	3	2	10	9	8	7	4	3	1	5	1	1	0	0
	5	4	3	11	9	6	1	6	2	3	2	1	1	1	0
	6	1	2	3	8	1	6	7	5	4	0	1	0	0	1
	7	4	5	9	2	6	4	1	4	2	2	0	1	0	0
	8	6	1	6	4	1	5	4	2	2	1	1	0	1	0
	9	1	2	1	2	3	6	3	2	1	3	1	1	0	1
	10	1	2	2	3	5	2	3	5	2	2	0	2	0	0
	11	0	2	0	0	1	3	4	2	3	2	0	1	0	0
	12	0	0	0	1	1	2	3	3	1	3	2	2	0	1
	13	0	0	0	0	1	1	0	0	3	0	4	3	3	1
	14	0	0	0	0	1	2	1	1	1	2	0	7	0	3
Number of alleles:		63	197	144	101	63	69	55	39	25	25	14	19	5	7

* This table shows there were 197 unique alleles (alleles found in a single progenitor only) among the SSP. Of those unique alleles: 144 were not found in any SSL; 18 were found in 1 of the SSL; 5 were present in 2 SSL, etc. Similarly, 7 alleles were present in each of the 13 SSP. Of those, 3 were subsequently present in all 14 SSL, while the others were found in 7, 9, 12, and 13 SSL.

† This column represents novel alleles (Labate et al., 1997). These are alleles that were not detected in any of the SSP, but were found in the SSL. In this case the number of novel alleles is overstated due to the missing data from 3 SSP.

Table 5. Markers that showed significant decreases in gene diversity in CBL and SSL. The table also shows the observed gene diversity, the probability of that observation (based on the simulation distributions), and the location and repeat length of the markers involved. Markers listed in bold are significant in both CBL and SSL. Markers are sorted by location.

Group	Marker	Observed gene diversity	Prob. of observed diversity	Marker location	Repeat length
SSL	umc1452	0.245	0.080	1.03	3
	umc1603	0.000	0.005	1.05	3
	umc1123	0.000	0.042	1.06	2
	umc1165	0.000	0.002	2.01	2
	umc1227	0.245	0.070	2.01	3
	umc1580	0.133	0.080	2.04	3
	bnlg1233	**0.463**	**0.060**	**2.08**	**2**
	mmc0381	**0.292**	**0.001**	**2.08**	**2**
	umc1049	0.000	0.001	2.08	2
	umc1135	0.000	0.003	3.07	3
	umc1844	0.000	0.096	3.08	2
	umc1062	0.272	0.040	3.09	3
	umc1813	0.000	0.063	3.09	3
	umc1550	0.245	0.040	4.03	2
	umc1869	0.133	0.010	4.06	3
	umc1058	0.133	0.008	4.11	2
	umc1624	0.000	0.006	5.04	3
	umc1502	0.245	0.080	5.05	2
	umc1153	0.245	0.025	5.09	3
	umc1002	**0.390**	**0.055**	**6.00**	**2**
	umc1229	0.391	0.010	6.01	2
	umc1187	0.133	0.080	6.05	3
	umc1388	0.133	0.080	6.05	3
	umc1788	0.000	0.080	7.00	5
	umc1567	0.245	0.040	7.03	3
	umc1412	0.337	0.100	7.04	3
	phi082	0.337	0.030	7.05	2
	umc1967	0.000	0.041	9.01	n/a
	umc1691	0.260	0.026	9.03	2
	umc1700	0.000	0.015	9.03	3
	umc1492	0.245	0.080	9.04	3
	umc1054	0.000	0.038	10.04	3
	umc1272	0.000	0.015	10.04	5

Table 5. Continued.

Group	Marker	Observed gene diversity	Prob. of observed diversity	Marker location	Repeat length
CBL	bnlg1175	0.375	0.048	2.04	2
	umc1875	0.000	0.018	2.06	2
	bnlg1233	**0.219**	**0.064**	**2.08**	**2**
	bnlg1746	0.000	0.040	2.08	2
	mmc0381	**0.219**	**0.066**	**2.08**	**2**
	umc1539	0.000	0.087	3.05	3
	umc1528	0.000	0.025	3.07	4
	umc1509	0.000	0.031	4.02	2
	umc1390	0.000	0.093	4.05	3
	umc1707	0.000	0.085	4.11	2
	umc1747	0.117	0.100	5.04	3
	phi126	0.375	0.056	6.00	2
	umc1002	**0.000**	**0.040**	**6.00**	**2**
	umc1483	0.000	0.065	8.01	3
	umc1121	0.000	0.015	8.05	4
	bnlg1012	0.357	0.100	9.04	2
	umc1120	0.000	0.077	9.04	5
	umc2017	0.000	0.086	10.03	3

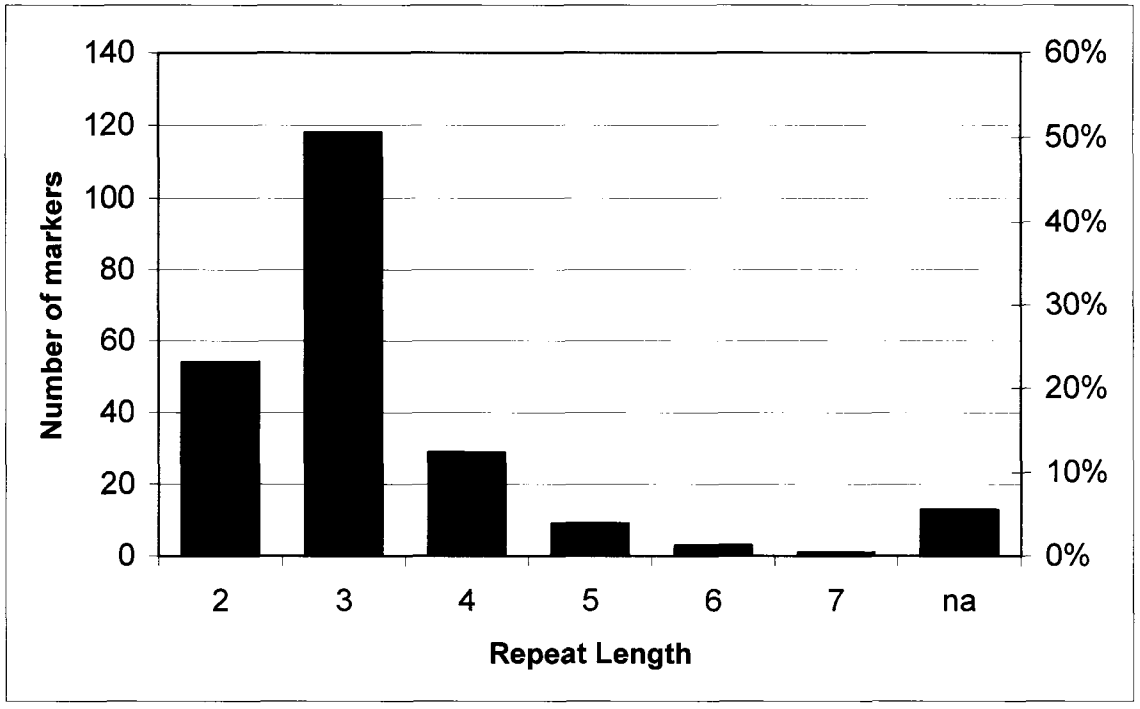


Fig. 1. Repeat length of the SSRs used in this study. Thirteen repeats had an unknown repeat length according the Maize GDB, and are represented by the "na" bar.

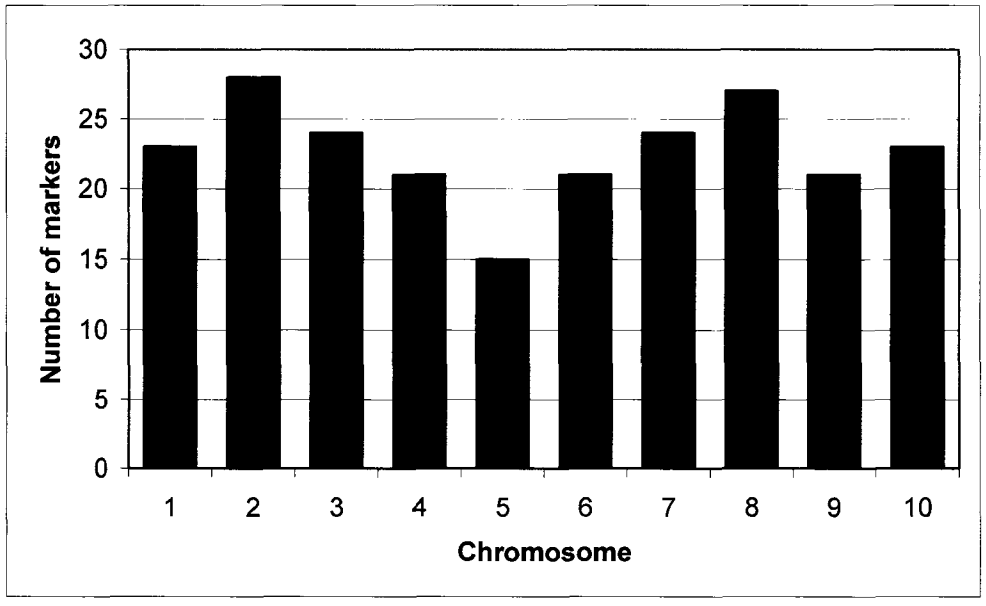


Fig. 2. Distribution of markers across chromosomes.

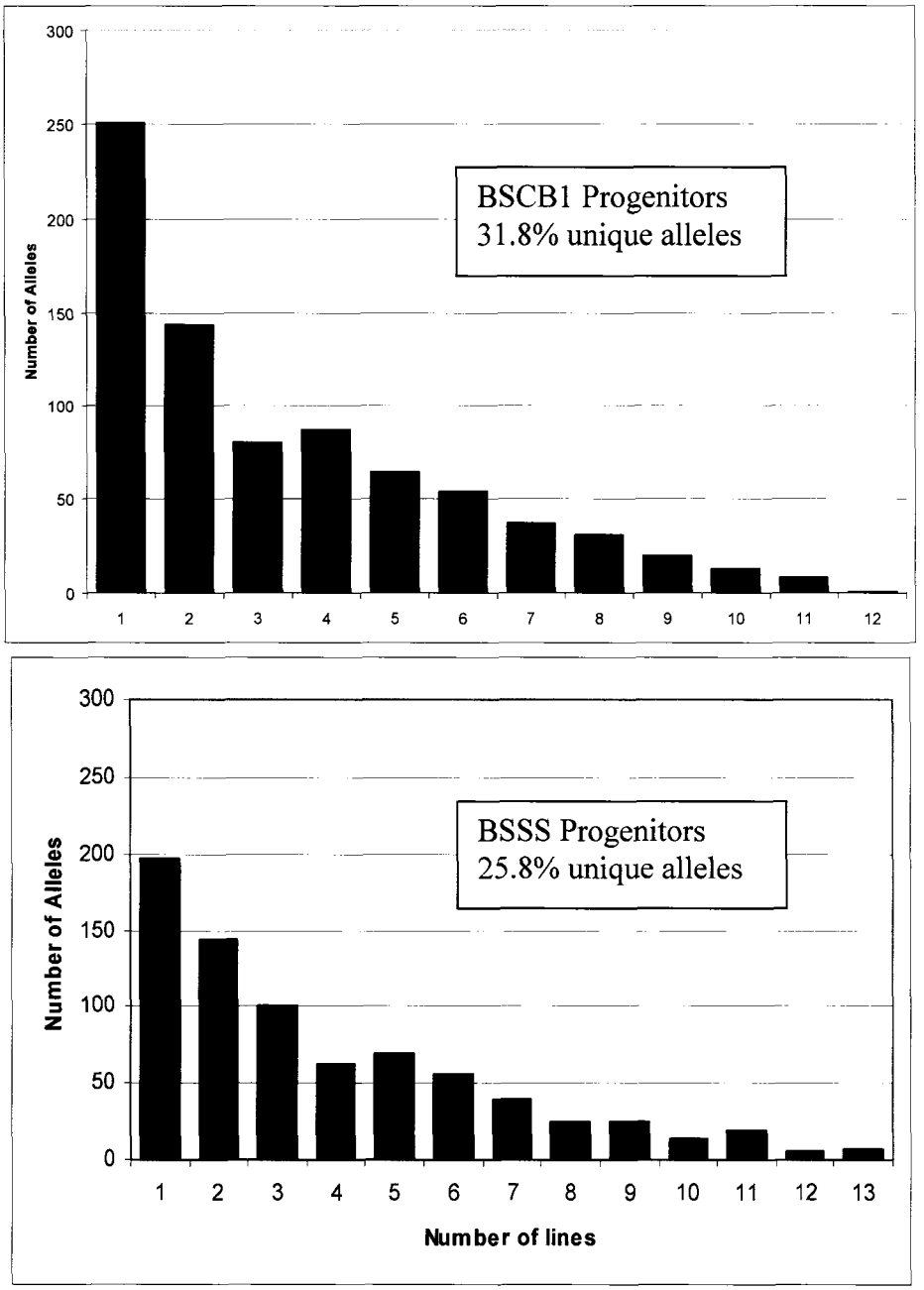


Fig. 3. Frequency distribution of the number of progenitor lines containing the same SSR allele for all loci in CBP and SSP.

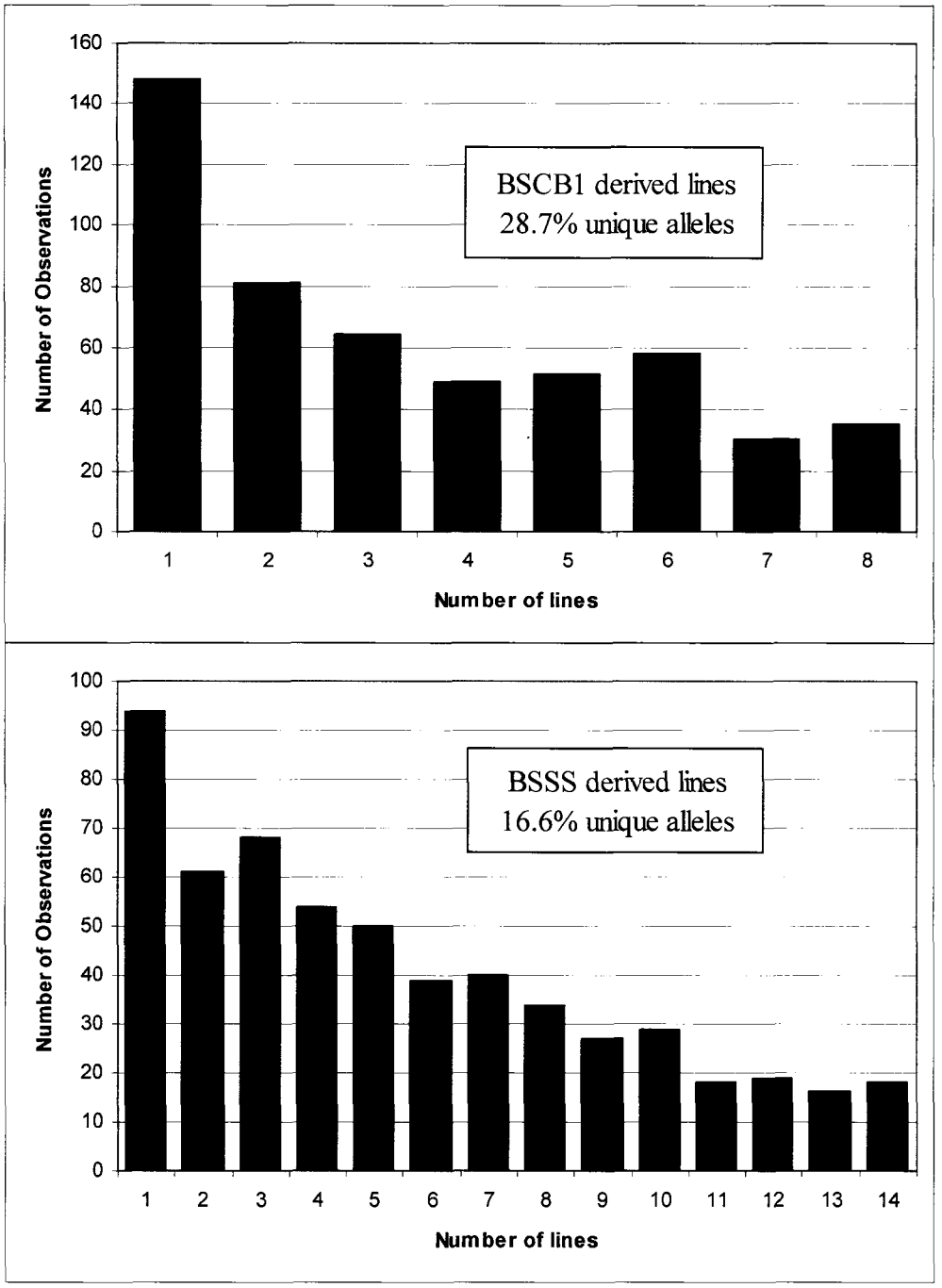


Fig. 4. Frequency distribution of the number of derived lines containing the same SSR allele for all loci in CBL and SSL.

CHAPTER 4. GENETIC DISTANCE AND PRINCIPAL COMPONENT ANALYSIS BETWEEN PROGENITORS AND DERIVED LINES OF TWO MAIZE POPULATIONS

A paper to be submitted for publication in *Crop Science*
Jim R. Rouse, Lori L. Hinze, Jode W. Edwards, and Kendall R. Lamkey

Abstract

Inbred progenitors of Iowa Stiff Stalk Synthetic (BSSS) and Iowa Corn Borer Synthetic #1 (BSCB1) maize populations, and inbred lines derived from each of the populations, were assayed for polymorphism at 227 microsatellite loci. The resulting genotypic information was used to calculate genetic distances between the progenitor and derived line germplasm groups.

Genetic distance between the progenitor groups was very low, indicating the progenitors were not highly divergent from each other. Rogers's distance (RD) between progenitor groups and derived line groups was identical in both BSSS and BSCB1, suggesting that the derived lines are equally divergent from their respective progenitor groups. The largest RD was between the two derived-line groups, about 35% greater than the progenitor-to-derived line distance. Rogers's distance between individual derived lines ranged from 15 to 54 in SSL and 25 to 56 in CBL, indicating some of the derived lines are very closely related.

The majority of the alleles in all germplasm groups in this study had very low frequencies, and approximately one-quarter to one-third of the alleles present in progenitor groups were not found in the derived lines.

Introduction

The Iowa Stiff Stalk Synthetic (BSSS) and Iowa Corn Borer Synthetic #1 (BSCB1) maize populations were developed by G.F. Sprague in the 1930s and 1940s. In 1949, shortly after Comstock et al. (1949) proposed reciprocal recurrent selection (RRS), Sprague initiated an RRS program with these two populations. The goal of RRS is to simultaneously improve both populations, while maintaining genetic variability within the populations. The interpopulation cross is the unit of evaluation in this program.

The RRS program between BSSS and BSCB1 is currently in the 16^{th} cycle of selection by Iowa's Cooperative Federal-State Maize Breeding Project. The longevity of this program and the successful concurrent inbred development program at Iowa State University have resulted in a prolific release of germplasm related to these two populations. Approximately 50 elite inbred lines related to these populations have been released to the public since the RRS program was initiated. Many of these lines have been used as parents in commercial hybrids or other synthetic populations (Zuber and Darrah, 1981; Darrah and Zuber, 1986; Rouse et al., 2003).

The success and popularity of the BSSS and BSCB1 populations has made them the subject of many phenotypic and molecular genetic studies over the past several decades. Selection progress has been described by Penny and Eberhart (1971), Martin and Hallauer (1980), Smith (1983), Keeratinijakal and Lamkey (1993a), Schnicker and Lamkey (1993), and Holthaus and Lamkey (1995).

According to Comstock et al. (1949), RRS would cause fixation of different alleles in the populations for traits controlled by overdominant loci. For traits under control of partial to complete dominance, RRS should cause an accumulation of the most favorable allele in

each population. The first scenario would cause an increase in genetic distance between the populations, while the second would lessen genetic distance (Hagdorn et al., 2003). At the same time, they acknowledge that genetic distance may also increase in cases of partial to complete dominance depending on the relative gene frequencies between the populations, and random genetic drift (Keeratinijakal and Lamkey, 1993b; Hanson and Moll, 1986).

Messmer et al. (1991) used RFLPs and allozymes to compare five elite inbreds from BSSS to the inbred progenitors of BSSS. This study was primarily a comparison of RFLPs versus allozymes for revealing pedigree relationships and grouping inbreds by genetic background, rather than an examination of the relationships between BSSS and BSSS-derived lines.

Our objective was to use SSR information to examine the genetic distance between the inbred progenitors of, and elite inbreds derived from, BSSS and BSCB1 maize populations. This information will be used to make conclusions about the genetic variability remaining in the BSSS and BSCB1 populations, and the genetic relationships of the inbreds derived from the populations.

Materials and Methods

We analyzed 47 maize inbred lines that we considered to be members of four separate groups. The first group, hereafter referred to as CBP, consisted of the twelve inbred progenitors of the Iowa Synthetic Corn Borer #1 population (BSCB1). The second group, hereafter referred to as SSP, consisted of 11 of the 16 progenitors of the Iowa Stiff Stalk

Synthetic populations (BSSS), and both parents of another progenitor.[†] Groups three and four, subsequently referred to as CBL and SSL, are comprised of eight inbred lines derived from BSCB1, and 14 inbred lines derived from BSSS, respectively. Table 1 contains a complete list of the inbreds used in this study and their origins. *Illinois Hy,* a progenitor of both populations, was assayed once at each locus, and the results were duplicated to be used in each progenitor group. Four of the BSSS lines did not originate from the RRS program, but were deemed to be related closely enough to be included as a BSSS-derived line (Hagdorn et al., 2003).

Microsatellite genotyping

All molecular data were collected at Cornell University in the Institute for Genomic Diversity. For details see Hinze (2003) or Hinze et al. (in review). Briefly, a 5 cm leaf section from ~2-week old plants was collected and freeze dried until DNA extraction. A miniprep protocol using CTAB extraction buffer (Mitchell et al., 1997) was then used to extract genomic DNA from the leaf samples, and the SSR regions were amplified using PCR and fluorescent-labeled primers. PCR products were separated by size on a fragment analyzer (Applied Biosystems, Foster City, CA) equipped with GENESCAN 3.1 software. GENOTYPER 2.1 software (Applied Biosystems) was used to identify markers and their specific alleles. If no amplification products were seen in the gel image, PCR was rerun for that entry/SSR combination.

[†] Of the original 16 BSSS progenitors, two (*CI617* and *F1B1*) have been lost. However the parents of *F1B1* (*Fe* and *IndB2*) are available, and when included with the other 14 extant BSSS progenitors thus constitute the 16 BSSS progenitors originally targeted for this study. Only 13 BSSS progenitors are included in this data set, though. Three lines, *I159*, *Ind. Tr 9-1-1-6*, and *A3G-3-1-3*, were inadvertently omitted from the laboratory procedures so no marker data are available on them.

Two hundred eighty-four SSR markers were chosen for analysis based on their distribution throughout the maize genome. Marker locations (bins) are based on information from the Maize GDB (Lawrence et al., 2004). A bin is a relative genetic map position on the chromosome that represents about 20 centiMorgans on the genetic map of maize (Hinze, 2003). Fifty-seven markers were discarded from the data set due to low PCR amplification, ambiguous results, or excess heterozygosity, leaving 227 markers to be analyzed for this study. The utilizable markers were distributed fairly evenly among the ten chromosomes, except for chromosome five, which had only 15 markers. Twenty-four percent of the SSRs were dinucleotide repeats, 52% were trinucleotide repeats, 13% were tetranucleotides. There were 13 repeats of unknown nucleotide length.

Genetic measurements

In this paper we present two genetic distance measurements, those of Rogers (1972) and Nei (1978), to compare our results with those of Hinze (2003), Labate et al. (1997), and Hagdorn (2003). Nei's distance was measured using the software PopGene (Yeh and Boyle, 1997), and Rogers's distance measurements were calculated using the software PowerMarker (Liu and Muse, 2003). We also multiplied the formula results by 100 simply to deal with integers.

For homozygous lines the Rogers's distance (RD) measurement corresponds with the proportion of marker alleles which differ between the two lines (Hagdorn et al., 2003). The formula for Rogers's distance (RD) is

$$\mathrm{RD} = \frac{1}{m}\sum_{i=1}^{m}\sqrt{\left(\frac{1}{2}\right)\sum_{j=1}^{a_i}\left(p_{ij}-q_{ij}\right)^2}\ , \text{ where}$$

m = number of loci;

a_i = number of alleles at the i^{th} locus;

p_{ij} , q_{ij} = allele frequencies of the j^{th} allele at the i^{th} locus in a pair or group of lines.

Nei (1972) defined genetic distance as:

$D = -\ln\left[J_{XY} / \sqrt{J_X J_Y} \right] = -\ln I$, where

X and Y represent two diploid populations;

$I = J_{XY} / \sqrt{J_X J_Y}$ = the normalized identity of alleles between X and Y over all loci,

where J_X, J_Y, and J_{XY} are the arithmetic means of the corresponding single-locus values;

$I_J = j_{XY} / \sqrt{j_X j_Y}$ = the normalized genetic identity of alleles between X and Y at a single locus;

$j_X = \sum x_i^2$ = the probability of identity of two randomly chosen alleles in population X;

$j_Y = \sum y_i^2$ = the probability of identity of two randomly chosen alleles in population Y;

$j_{XY} = \sum x_i y_i$ = the probability of identity of two alleles, one from X and one from Y;

x_i and y_i, = the frequencies of the i^{th} alleles in X and Y.

With a modification to overcome limitations due to small sample sizes, Nei (1978) modified the genetic distance measurement to

$ND = -\ln\left[\hat{G}_{XY} / \sqrt{\hat{G}_X \hat{G}_Y} \right]$, where

$\hat{G}_X$ and $\hat{G}_Y$ = the multilocus averages of $(2n_x J_X - 1)/(2n_x - 1)$ and $(2n_Y J_Y - 1)/(2n_Y - 1)$, and

$\hat{G}_{XY} = J_{XY}$.

Results and Discussion

Without regard to germplasm groups, 20% of the alleles have a frequency of 0.02 or less, 50% have a frequency of 0.15 or less, and 70% have a frequency of 0.30 or less (Fig. 1). These results are not particularly surprising since the frequency of most alleles within a group was relatively low. For example, nearly 32% of the alleles found in CBP were unique to a single progenitor, another 18% were found in only two progenitors, and 10% in three progenitors. This gives a cumulative total of 60% of the alleles in CBP with a frequency less than 0.25 (three progenitors out of 12, or six alleles out of 24 total alleles). Results are similar in SSP, with nearly 25% of the alleles unique to a single progenitor, 19% in two progenitors, and 13% in three progenitors. As an average across all groups, 63% of the alleles were found in four or fewer lines within each group. This indicates that most of the alleles would have low to very low frequencies.

Figures 2 and 3 compare the allele frequencies between the progenitor groups and their respective derived-line groups. The progenitor groups had significantly more low-frequency alleles than the derived-line groups in both BSSS and BSCB1. The SSL contained about 25% fewer alleles than the SSP, and the CBL about 34% fewer alleles than the CBP (data not shown).

Genetic distance

The proportion of alleles in common among the germplasm groups was determined using genetic distance measurements. For comparison with previous studies we calculated both Rogers's distance (RD) and Nei's unbiased genetic distance (ND) (Rogers, 1972; Nei, 1978). The ND between the CBP and SSP was 9 (Table 2), similar to previous results of 8 (Hinze, 2003), and 7 (Labate et al., 1997). Thus, the two progenitor groups are very closely related. The ND between the CBP and CBL is 18, and the distance between the SSP and SSL is 19. Since Hinze (2003) and Labate et al. (1997) each used the populations per se, no other genetic distances are directly comparable between this study and their studies. The results of the ND and RD calculations support similar conclusions, so the remaining discussion will focus on RD for simplicity.

The RD between CBP and SSP was 20, compared to RD of 19 found by Hagdorn et al. (2003). This supports previous conclusions that the progenitor inbreds are not highly divergent. However, the RD between the derived line groups was 38, nearly twice the distance between the progenitor groups. Hagdorn et al. (2003) made similar comparisons, but the derived lines in their study were separated into those derived from C0 and those derived from advanced cycles. In their study, the RD between the advanced cycle lines was 50, somewhat higher than the distance in our study. To make this a more acceptable comparison, we should separate our derived lines in a similar manner. We are currently in the process of collecting SSR information on several more derived inbred lines from both populations, and will wait to do this comparison until more data are collected. Instead, we will compare the RD between derived lines and the progenitors, as well as between derived lines and their progenitor group.

The individual line-by-progenitor RDs ranged from 48 to 68 among the BSSS inbreds(Table 3), and 46 to 66 among the BSCB1 inbreds (Table 4). In both tables, the derived lines are listed in the order of cycle of origin, with those from early cycles on the left side of the table. The mean RDs are as described in the table notes, simply the row or column means. We suspected there may be a trend to greater distances as one moves left to right across the table, but this appears to be only marginally so.

The mean RD of individual SSL from the SSP group ranged from 49 to 52, with a mean of 51 (Table 3). These distances are not the same as the mean of the line-by-progenitor RD. In all cases the distance from the SSP group is less than the mean of the distances from the individual progenitors. This is likely due to the fact that frequencies in individual lines are either 0 or 1 (barring heterozygotes), while the SSP group may may contain several alleles with low frequencies. This would have the effect of making the group appear closer than the mean of the individual lines.

The mean RD of individual CBP from the CBP group ranged from 48 to 51, with a mean of 49 (Table 4). As with BSSS (Table 3), the RD between the derived lines and the progenitor group is less than the mean of the individual line-by-progenitor RD.

Rogers's distances between the BSSS derived lines ranged from 15 to 54 (Table 5). The bottom row of this table contains the same data as the bottom row of Table 3 for easy comparison of the line-by-line distance with the line-by-SSP distance. Most of the line-by-line distances are a bit less than the line-by-progenitor distances, but there are a few distances that are significantly lower. The following pairs of inbreds have the smallest RDs in the table, and thus appear to be much more closely related to each other than to the rest of the lines (RD in parentheses): *B64* and *B68* (15); *B105* and *B111*, *B104* and *B112* (both 24); *B94*

and *B105*, *B94* and *B111* (both 28). In all cases, the mean line-by-line RD is less than the line-by-CBP RD.

The line-by-line RDs for the BSCB1 inbreds range from 25 to 56 (Table 6). Again, the line-by-CBP distance is shown for comparison. Of particular note here are the relatively low distances between the later lines. As a reminder, the first two inbreds are derived from BSCB1 C0, and the remainder from C7 though C11. Of all possible distances between the six advanced-cycle lines, all are less than 40, most are closer to 30, and the *B99*-to-*B112* distance is only 25. In contrast, the individual progenitor-by-progenitor distances range from 52 to 67 (data not shown).

The two C0 inbreds (*B42* and *B54*) have a larger mean RD from other CBL than from the CBP as a group. All the late-derived CBL have mean RDs substantially less than the mean RD of the two C0-derived inbreds, and considerably less than the line-by-CBP RD. The CBL, particularly the advanced-cycle lines, are much more closely related to each other than the CBP are to each other.

Principal component analysis (PCA)

The first three PCs describe 8.4, 5.7, and 5.1% of the total variation among the inbreds, respectively, for a cumulative total of 19.2% (Fig. 4). While this is a small fraction of the total variability, there is a great deal of separation between the groups in Fig. 3. Using GGobi (www.GGobi.org) to compare various projections of the data with varying numbers of PC axes, Fig. 3 shows as much separation as the projections using more PCs. Adding more PCs tended to make the germplasm groups less distinct visually, (likely caused by attempting to display multidimensional images in a 2-dimensional medium).

There are some observations that can be made from the 3-D projections that are not apparent from the static, 2-D image of Fig. 4. For example, the overall distribution of the data resembles the shape of a saddle, with the CBP and SSP making up the 'seat', with the 'legs' consisting of the CBL and SSL. Furthermore, even though the progenitor inbreds form an amorphous, well-spaced cluster, there is some separation of the progenitors that follows population subdivision, in that the BSSS and BSCB1 progenitors appeared to remain on their own half of the saddle seat. Little has been said about this separation of the progenitors, though Messmer et al. (1991) noticed that progenitors of non-Reid Yellow Dent (RYD) origin seemed to be more distantly related than those of RYD origin.

Although we do not have population data, we do have some inbred lines derived from the C0 of each population. The two CBL derived from C0 appear to belong in the progenitor cluster, which is not a surprising result. In the case of BSSS, though, the C0 lines are separate from the progenitors. This separation has been previously noted in the C0 of the populations—from each other and from their respective progenitors (Hinze, 2003; Labate et al., 1997; 1999). The separation between the progenitors and their C0 population was presumably due to the random mating generations that occurred after intermating the progenitors and before the C0 was declared formed. The BSSS C0 was genetically more distant from its progenitors than the BSCB1 C0 from its progenitors, possibly because of more generations of random mating in the BSSS than the BSCB1 (Labate et al., 1997; 1999).

Certain pairs of the progenitors appear to be very closely related, as they remain together throughout all GGobi projections of the data. These pairs include inbreds *I224* and *Ind 461-3*, *Oh3167B* and *Le23*, *Oh07* and *Illinois Hy* (*Hy* was used as a progenitor in both populations). Based on the pedigree information from Table 1 there is no readily apparent

relationship between the inbreds within these pairs. Upon further inspection of the raw data, there has almost certainly been a coding error with the data for *I224* and *Ind 461-3*, and with the data for *Oh3167B* and *Le23*. These inbreds will be reassayed, and the new data reanalyzed, before publishing these articles.

Some of the progenitors appear to be individual outliers, located graphically either between the bulk of the progenitor cluster and the cluster of derived lines, or appearing somewhat equidistant from the derived line groups, but somewhat separate from the progenitor cluster as well (for example, in the figurative position of the saddle horn). These lines are *Oh40B*, *L317*, and *CC5* in BSCB1, and *WD456*, *Illinois 12E*, *CI540*, and *IndB2* in BSSS.

Six of the eight CBL form a loose cluster separate from the entire progenitor group. Two of the CBL, though, remain firmly entrenched in the center of the progenitor cluster. These inbreds, *B42* and *B54*, were both derived from BSCB1 C0, perhaps accounting for their close relationship with the progenitors.

The SSL are approximately the same distance from the progenitor cluster as the CBL, but the SSL remain in a very tight cluster in all projections. Two inbreds, *B101* and *B37*, seem to stray a little closer to the progenitor cluster at times, but never too far from the rest of the SSL. These two inbreds were also derived from the C0 population, but three other inbreds in SSL also originated in BSSS C0, so there is no obvious explanation for the differences between these lines.

Conclusions

The goals of reciprocal recurrent selection are to increase the population mean and maintain genetic variability (Comstock et al., 1949). We have made several references to authors measuring the progress of selection in the populations per se. But how do we know if we are maintaining genetic variability? One indication might be the continued success of an inbred development program. Based on the results of the genetic distance measurements, though, that may not always be the case. The very low genetic distances between the several pairs of inbred lines mentioned above indicate that some of the inbred lines that have been released may not be all that different from each other.

The information gained from Fig. 4 is useful in the context of heterotic groups. We did not spend much time discussing the heterotic groups in this paper, and do not intend to start doing so now. But the clear separation of the derived line groups from the progenitor groups lends evidence to the idea that heterotic groups can be created by the actions of breeders, and are not necessarily inherent in the germplasm.

One experiment that would provide an interesting supplement to the data presented here would be to observe these inbred lines crossed onto a few common testers. In addition to attempting to detect significant phenotypic differences between the inbreds, the differences in phenotype could possibly be correlated with the genetic distances. If there are no differences between some of the advanced-cycle inbreds, serious consideration will have to be devoted to continuing the RRS program. This would depend upon the outcome of the testcross evaluations, of course, because the genetic distances by themselves provide no information regarding the performance of these inbred lines.

References

Comstock, R.E., H.F. Robinson, and P.H. Harvey. 1949. A breeding procedure designed to make maximum use of both general and specific combining ability. Agronomy Journal. 41:360-367.

Darrah, L.L., and M.S. Zuber. 1986. 1985 United States farm maize germplasm base and commercial breeding strategies. Crop Sci. 26:1109-1113.

Hagdorn, S., K.R. Lamkey, M. Frisch, P.E.O. Guimaraes, and A.E. Melchinger. 2003. Molecular genetic diversity among progenitors and derived elite lines of BSSS and BSCB1 maize populations. Crop Sci. 43:474-482.

Hanson, E.D., and R.H. Moll. 1986. An analysis of changes in dominance-associated gene effects under intrapopulation and interpopulation selection in maize. Crop Sci. 26:268-273.

Hinze, L.L. 2003. Molecular marker analysis of population genetic structure and progress from reciprocal recurrent selection in two Iowa maize (*Zea mays* L.) populations. PhD dissertation. Iowa State University, Ames, IA.

Hinze, L.L., S. Kresovich, J. Nason, and K.R. Lamkey. (In review.) Population genetic diversity in a maize reciprocal recurrent selection program. Crop Sci.

Holthaus, J.F., and K.R. Lamkey. 1995. Response to selection and changes in genetic parameters for 13 plant and ear traits in two maize recurrent selection programs. Maydica. 40:357-370.

Keeratinijakal, V., and K.R. Lamkey. 1993a. Responses to reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 33:73-77.

Keeratinijakal, V., and K.R. Lamkey. 1993b. Genetic effects associated with reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 33:78-82.

Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 1997. Molecular genetic diversity after reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 37:416-423.

Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 1999. Temporal changes in allele frequencies in two reciprocally selected maize populations. Theor. Appl. Genet. 99:1166-1178.

Lawrence, C.J., Q. Dong, M.L. Polacco, T.E. Seigfried, and V. Brendel. 2004. MaizeGDB, the community database for maize genetics and genomics. (Available at http://maizegdb.org). Nucleic Acids Res. 32:D393-D397.

Liu, K., and S. Muse. 2003. PowerMarker: New genetic data analysis software. Ver. 3.0. Free program distributed by the authors over the internet from http://www.powermarker.net.

Martin, J.M., and A.R. Hallauer. 1980. Seven cycles of reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 20:599-603.

Messmer, M.M., A.E. Melchinger, M. Lee, W.L. Woodman, E.A. Lee, and K.R. Lamkey. 1991. Genetic diversity among progenitors and elite lines from the Iowa Stiff Stalk Synthetic (BSSS) maize population: comparison of allozyme and RFLP data. Theor. Appl. Genet. 83:97-107.

Mitchell, S.E., S. Kresovich, C.A. Jester, C.J. Hernandez, and A.K. Szewc-McFadden. 1997. Application of multiplex PCR and fluorescence-based, seme-automated allele sizing technology for genotyping plant genetic resources. Crop Sci. 37:617-624.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics. 89:583-590.

Penny, L.H., and S.A. Eberhart. 1971. Twenty years of reciprocal recurrent selection with two synthetic varieties of maize (*Zea mays* L.). Crop Sci. 11:900-903.

Rogers, J.S. 1972. Measures of genetic similarity and genetic distance. Studies in genetics. VII. Univ. Texas Publ. 7213:145-153.

Rouse, J.R., L.L. Hinze, and K.R. Lamkey. 2003. Selection history and developments from the BSSS and BSCB1 maize populations. *In* International Symposium on Plant Breeding poster presentations [CD-ROM]. CIMMYT, Mexico City, Mexico.

Schnicker, B.J., and K.R. Lamkey. 1993. Interpopulation genetic variance after reciprocal recurrent selection in BSSS and BSCB1 maize populations. Crop Sci. 33:90-95.

Smith, O.S. 1983. Evaluation of recurrent selection in BSSS, BSCB1, and BS13 maize populations. Crop Sci. 23:35-40.

Yeh, F.C., and T.J.B Boyle. 1997. Population genetic analysis of co-dominant and dominant markers and quantitative traits. Belgian Journal of Botany 129:157.

Zuber, M.S., and L.L. Darrah. 1981. 1979 U.S. corn germplasm base. Proc. 35th Ann. Corn & Sorghum Res. Conf. ASTA. pp. 234-249.

Table 1. Inbred lines analyzed in this study.

Line	Origin/Pedigree
	Progenitor lines of Iowa Corn Borer Synthetic #1 (CBP) ‡
A340	4-29 X 64
CC5	Golden Glow (renamed W23)
I205	Iodent
Ill. Hy	Illinois High Yield
K230	No origin listed, possibly from Midland Yellow Dent
L317	LSC
Oh07	CI.540 X Ill. L
Oh33	Clarage
Oh40B	Composite of 8 LSC lines
Oh51A	[(Oh56 X Wf9) Oh56]
P8	Palin Reid
R4	Funk Yellow Dent
	Progenitor lines of Iowa Stiff Stalk Synthetic (SSP) ‡
CI.187-2	Krug-Nebraska Reid strain X Iowa Gold Mine
CI.540	Illinois 2-ear Variety
I224	Iodent
Ill. 12E	Source unknown
Ill. Hy	Illinois High Yield
Ind. 461-3	Reid Medium (Duddleston No. 461)
Ind. AH83	Funks 176A
Ind. B2	Troyer Reid (Late Butler)
Ind. Fe	Troyer Reid (Early)
LE23	Illinois Low Ear
Oh3167B	Echelberger Clarage
Os420	Osterland Yellow Dent
WD456	Walden Dent

‡ All progenitor pedigrees are listed and referenced in Hagdorn et al. (2003).

Table 1. Continued.

Line	Origin/Pedigree
Elite lines derived from Iowa Corn Borer Synthetic #1 (CBL)	
B42	BSCB1C0
B54	BSCB1C0
B90	BSCB1(R)C7
B91	BSCB1(R)C8
B95	BSCB1(R)C7
B97	BSCB1(R)C9
B99	BSCB1(R)C10
B112	BSCB1(R)C11
Elite lines derived from Iowa Stiff Stalk Synthetic (SSL)	
B101	BSSSC0
B37	BSSSC0
B67	BSSSC0
B68	BSSSC0
B69	BSSSC0
B64	BSSS(R)C4
B89	BSSS(R)C7
B94	BSSS(R)C8
B105	BSSS(R)C9
B111	BSSS(R)C9
B73	BSSS(HT)C5
B84	BS13(S)C0
B104	BS13(S)C5
B110	BS13(S)C5
BSSS progenitor lines that were omitted from this study	
I159	Iodent
Ind Tr 9-1-1-6	Reid Early Dent (Troyer strain)
A3G-3-1-3	[BL345B X IAI129] (both strains of RYD)

Table 2. Rogers's (1972) genetic distance (x 100) and Nei's (1978) unbiased genetic distance (x 100) between four germplasm groups.*

Rogers's distance				Nei's distance			
	CBL	**CBP**	**SSL**		**CBL**	**CBP**	**SSL**
CBP	28			**CBP**	18		
SSL	38	31		**SSL**	37	23	
SSP	32	20	28	**SSP**	24	9	19

* SSP = BSSS progenitors; CBP = BSCB1 progenitors; SSL = BSSS derived lines; CBL = BSCB1 derived lines

Table 3. Rogers's distances (x 100) between individual SSL (column headers) and SSP inbreds (row headers), and between individual SSL and the SSP group (bottom row).

	B101	B37	B67	B68	B69	B64	B73	B84	B89	B94	B105	B111	B104	B110	Mean[†]
I224	56	57	53	53	52	51	57	57	56	57	51	57	57	57	55
Os420	54	57	57	59	58	62	60	57	53	48	48	55	55	60	56
WD456	54	54	50	48	55	49	57	54	50	51	52	52	57	60	53
Ind. 461-3	56	58	54	53	54	53	57	58	56	58	52	58	57	58	56
Ill. 12E	57	60	60	63	60	59	61	57	59	60	60	59	59	59	59
CI540	63	65	63	58	64	61	67	63	66	61	62	63	64	65	63
Hy	53	50	57	56	57	55	53	54	52	51	53	53	59	58	54
Oh3167B	53	60	57	57	58	54	57	53	57	55	57	60	54	54	56
AH83	52	56	53	56	54	57	57	56	58	52	49	57	61	58	55
CI187-2	57	61	56	63	55	67	67	65	59	53	58	59	68	62	61
Le23	53	60	58	57	58	54	57	53	57	56	58	59	54	55	56
IndB2	61	59	60	59	57	57	57	53	57	58	55	57	56	60	58
Fe	57	53	57	54	58	55	57	55	60	59	57	60	61	57	57
Mean[‡]	56	58	56	57	57	57	59	56	57	55	55	58	59	59	57
SSP Group[*]	50	51	50	50	50	50	52	50	50	49	49	51	52	52	51

[†] This is the mean RD of all SSL from a specific progenitor.
[‡] This is the mean RD of all SSP from a specific derived line.
[*] This is the RD of a specific line from the SSP as a group.

Table 4. Rogers's distances (x 100) between individual CBL (column headers) and CBP inbreds (row headers), and between individual CBL and the CBP group (bottom row).

	B42	B54	B90	B95	B91	B97	B99	B112	Mean[†]
R4	54	58	54	62	53	54	59	55	56
K230	54	52	60	54	63	59	61	60	58
I205	55	55	53	53	57	56	55	53	55
Oh40B	52	58	60	55	63	62	63	62	59
Hy	46	46	58	58	62	59	60	56	56
Oh07	54	57	59	62	62	59	66	65	60
Oh33	57	49	60	60	56	58	61	58	58
Oh51A	61	57	53	57	57	54	54	58	56
P8	52	48	50	53	54	49	56	49	51
L317	60	60	49	48	48	51	51	50	52
CC5	59	57	51	60	62	58	55	52	57
A340	55	58	52	52	58	52	57	53	55
Mean[‡]	55	55	55	56	58	56	58	56	56
CBP Group[*]	48	48	48	49	51	49	51	49	49

[†] This is the mean RD of all SSL from a specific progenitor.
[‡] This is the mean RD of all SSP from a specific derived line.
[*] This is the RD of a specific line from the SSP as a group.

Table 5. Rogers's distances (x 100) between inbred lines derived from BSSS, and the Rogers's distances between the derived lines and the BSSS progenitor group.

	B101	B37	B64	B67	B68	B69	B73	B89	B94	B105	B111	B84	B104	B110
B37	50													
B64	51	54												
B67	40	46	43											
B68	46	53	15	41										
B69	46	54	42	37	39									
B73	47	47	47	52	40	42								
B89	43	48	54	45	53	45	52							
B94	47	51	50	41	48	40	46	36						
B105	43	51	46	38	44	41	42	37	28					
B111	47	52	52	45	48	45	51	32	28	24				
B84	47	44	43	46	46	45	31	45	45	42	50			
B104	48	52	42	48	40	37	37	49	51	44	45	37		
B110	42	51	48	45	46	40	39	42	48	42	48	36	24	
Mean	46	50	45	44	43	43	44	45	43	40	43	43	43	43
SSP Group	50	51	50	50	50	50	52	50	49	49	51	50	52	52

Table 6. Rogers's distances (x 100) between inbred lines derived from BSCB1, and the Rogers's distances between the derived lines and the BSCB1 progenitor group.

	B42	B54	B90	B91	B95	B97	B99	B112
B54	55							
B90	54	45						
B91	53	52	32					
B95	56	49	35	39				
B97	55	48	29	34	35			
B99	56	53	34	30	38	32		
B112	55	49	32	31	35	30	25	
Mean	55	50	38	39	41	38	38	37
CBP Group	48	48	48	51	49	49	51	49

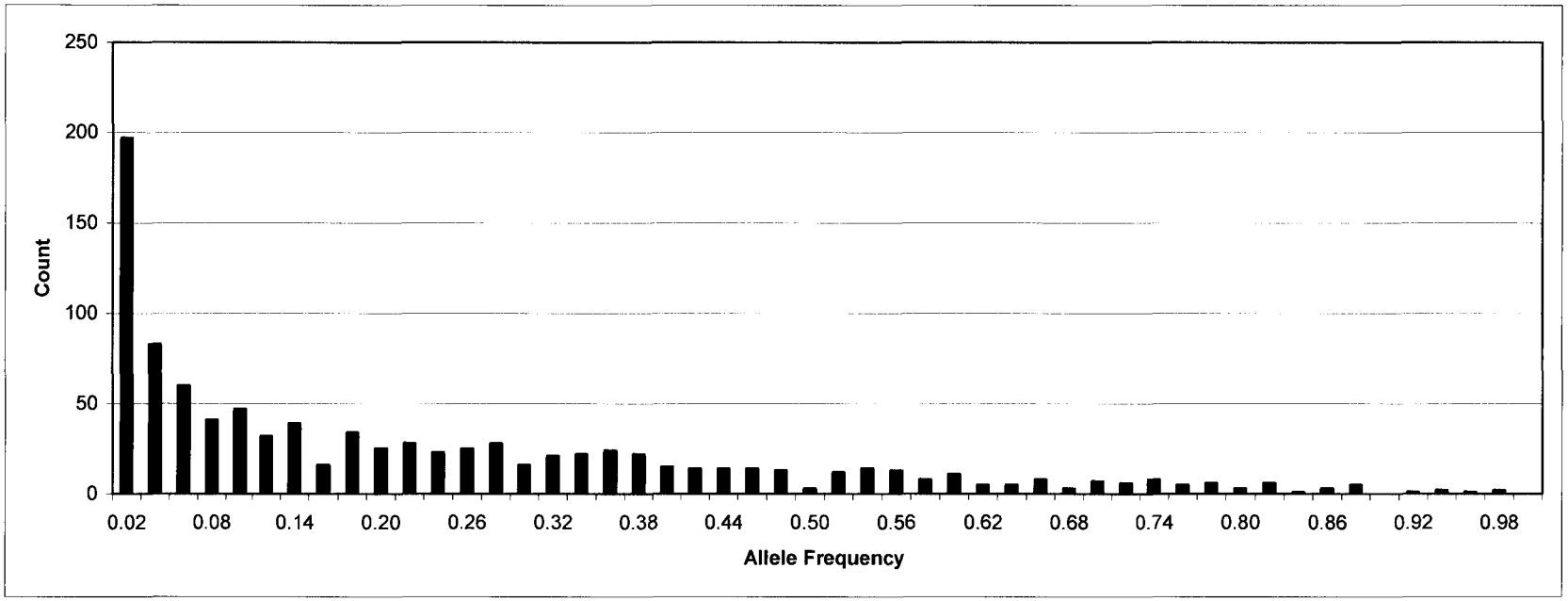


Fig. 1. Histogram of allele frequencies of all alleles from all germplasm groups. Twenty percent of alleles have a frequency of 0.02 or less; 50% have a frequency of 0.15 or less, and 70% are at 0.3 or less.

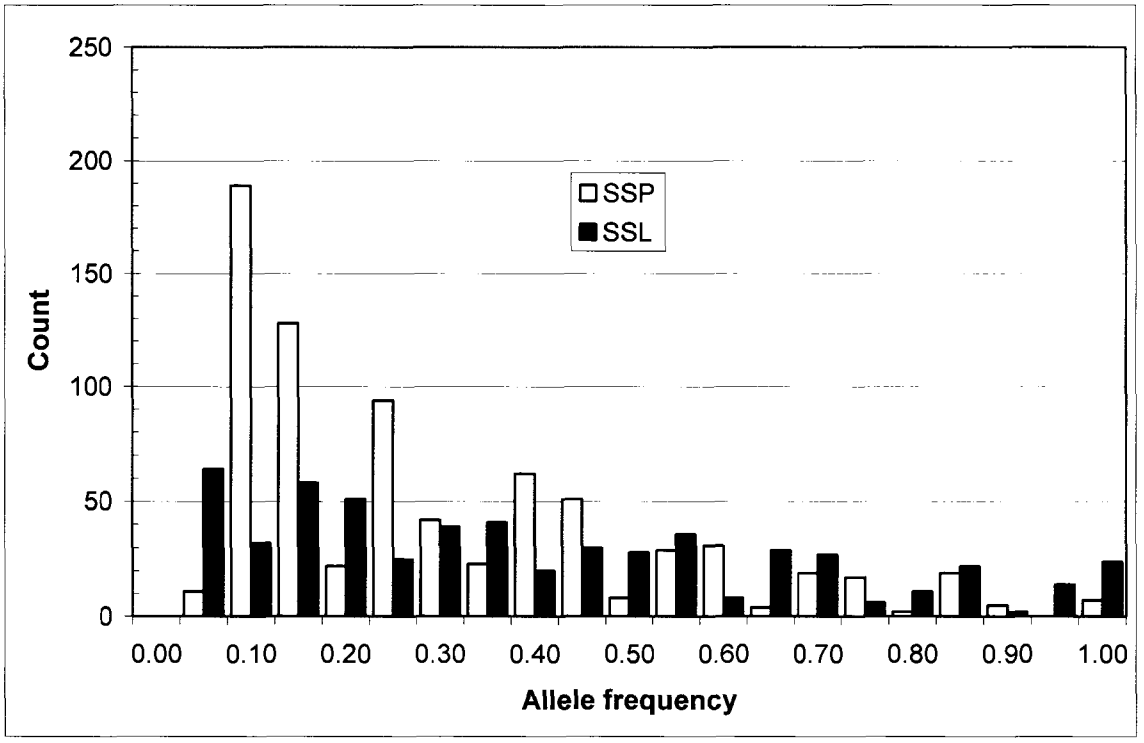


Fig. 2. Allele frequencies within the SSP and SSL germplasm groups.

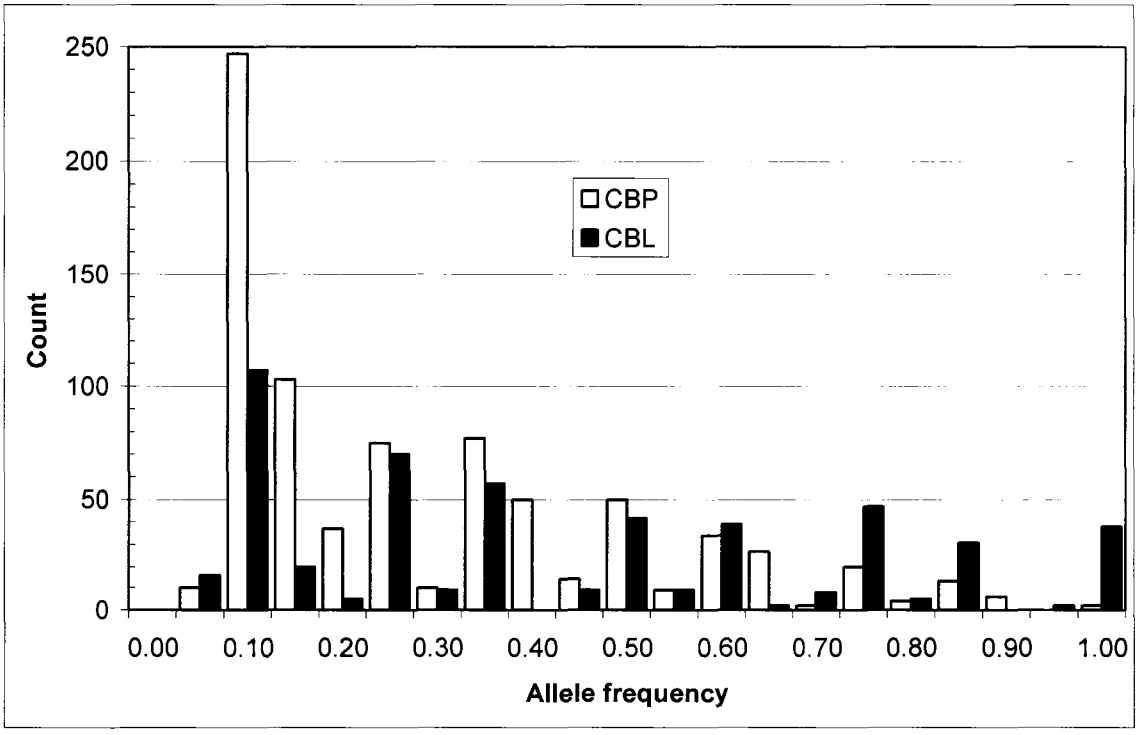


Fig. 3. Allele frequencies within the CBP and CBL germplasm groups.

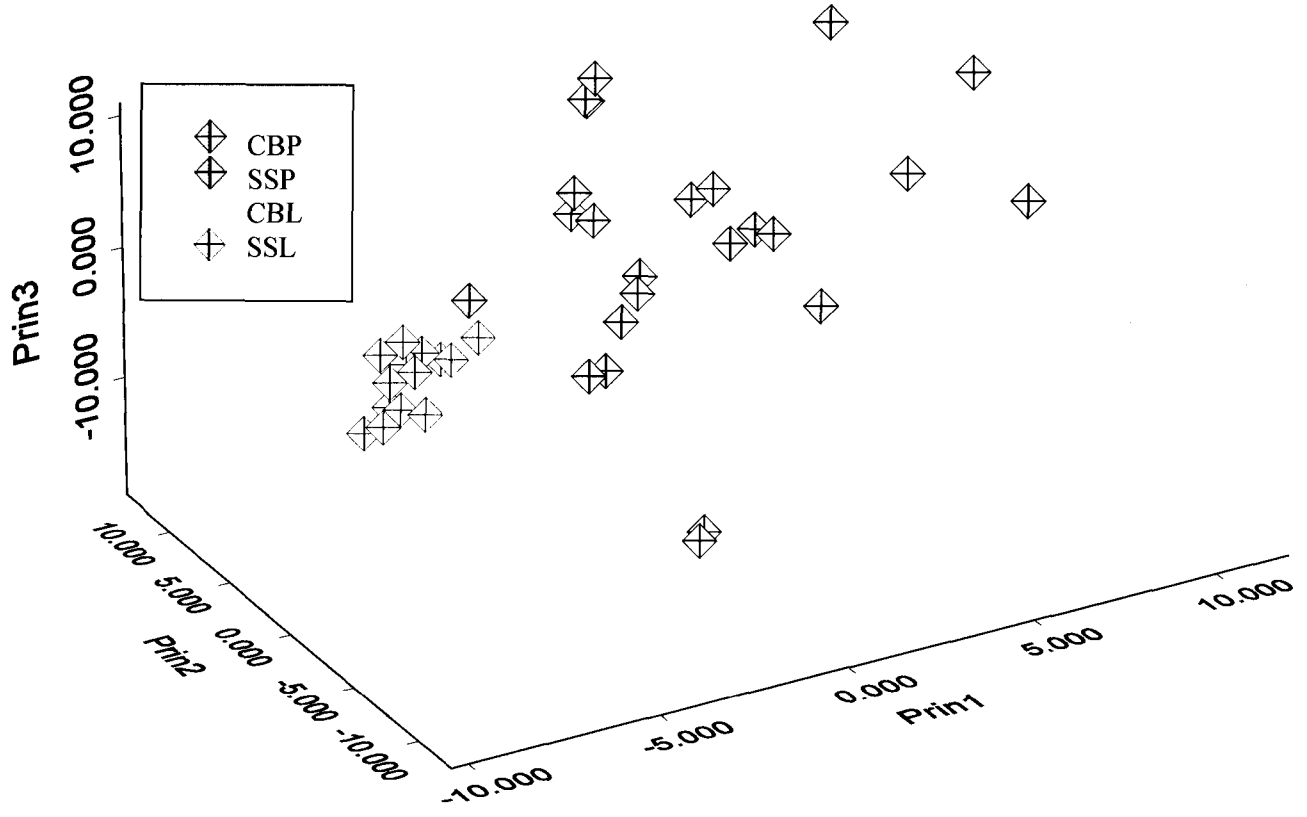


Fig. 4. Diagram of the inbreds used in this study plotted by the first three principal components. The inbreds are grouped by color into their respective germplasm groups.

CHAPTER 5. GENERAL CONCLUSIONS

In these studies, we are interested in genetic changes from a breeding perspective, but we are using population biology methods to determine what those changes are. These studies are unlike a normal testing of parentage, with marker data on diploid progeny and on one or both of the possible parents. Considering number of progenitors and the many cycles of selection and recombination in the RRS program, the derived inbred lines in this study have multiple possible contributions by multiple possible parents. Few unique alleles from the progenitors appear in the derived lines, which makes determining which progenitor contributed a given allele difficult.

We examined allele frequencies, numbers of alleles, gene diversity, and genetic distances to determine if the various germplasm groups can be distinguished by these methods. By using simulations we differentiated among selection, drift, or some combination of the two processes as the forces acting on the alleles. Even with simulations, we cannot be sure if our observed results are due to drift or selection. For example, a unique allele among the progenitors may have very little effect, and hence go unnoticed in the resulting population. But if that once-unique allele shows up in one or two of the selections used to advance the population cycle, it now has a much larger frequency and possibly a much larger effect, one that may be measurable. But is this an example of selection or drift?

Conceptually, it seems that if the unique allele from the progenitor was found in a selected individual because of it's favorable effect on some measurable trait that we selected for, then we can rightly conclude that we have changed that allele frequency due to selection. Conversely, if that allele is neutral or has a favorable, albeit unmeasurable, effect on a

selected trait in the population, it seems that resultant changes in that allele's frequency must be considered drift, not selection. But what if the higher allele frequencies in the advanced cycles of the population result in a measurable effect? Is it now considered selection, even though it was permitted only by the prior providence of random genetic drift?

Perhaps the distinction between selection versus drift gains more importance in the context of the populations under study, and the traits for which they are undergoing improvement. We have discussed the analysis of an allele with a measurable effect on a selectable trait. Breeders have spent years using various methods trying to improve populations for an assortment of phenotypic values. A common method of estimating progress due to selection is to compare an improved cycle of the population with the original C0 population in a field trial. Of course, the improved population is considered to be superior due to successful breeding and selection methods.

But we cannot rule out that progress has been due to drift rather than selection. One of the major obstacles of these types of studies is that very few recurrent selection programs are replicated with unselected lines. To be able to positively declare that progress is due to selection, we should compare the advanced selection cycles with several cycles that have used unselected (random) lines to form the advanced cycles. This is, of course, impractical for any conceivable inbred line development program.

With the advent of molecular genetic techniques, breeders began trying to assess molecular genetic changes as a cause for, or an effect of, the improvement due to selection. At the same time, there is speculation that molecular genetics might shed light on the gene action at QTL, as discussed in the Introduction of Chapter 4. But these issues raise additional dilemmas. In a single paragraph of Chapter 4, we discussed the hopes of using genetic

distance to establish gene action, and then spoiled that hope by illustrating situations that might confound results. There are also problems associating molecular changes with phenotypic selection. In the studies described in this dissertation, and most of those referenced herein, the emphasis has centered on detecting molecular genetic changes that seem to fall outside the scope of random genetic drift. The key component that is missing, and rarely seems to accompany these studies, is to continue to scrutinize the populations and try to relate the molecular changes to changes of agronomic importance.

An enigma that hinders progress in this area is that there are several good theoretical reasons to doubt that a strong connection exists between levels of molecular and quantitative genetic diversity within populations (Lynch, 1996):

1. There is a great disparity between the rate of mutation for molecular markers (around 10^{-8} to 10^{-5} per generation) versus the rate of introduction of quantitative trait variability (about 10^{-3} to 10^{-2} per generation). The result is that populations that have undergone severe bottlenecks will exhibit this evidence in their genetic signature for tens of thousands of years, even though the populations have had ample time to recover normal levels of heritable variation.
2. In small populations, the variation in quantitative genetic parameters can be quite large, even for quantitative genetic traits with purely additive gene action. Also, estimates of the quantitative genetic parameters are subject to substantial sampling error.
3. The sampling variance of measures of molecular variation can be very high. Even when the survey sample (numbers of individuals and markers) is large enough to

> allow confidence in the molecular assessment, the marker loci provide little insight into the conditions at loci underlying adaptive [or selective] variation.

Our study has highlighted some areas of the genome that appear to have undergone changes due to forces other than selection. Some of the areas we identified overlap with areas recognized by Hinze (2003) and Labate et al. (1999). It seems reasonable to assume that if we are to discover alleles with a favorable effect on one of our selectable traits (yield, moisture, or standability), we should begin association mapping studies concentrating on these areas first. Increased efforts to identify additional useful markers, or to locate candidate genes, should increase the chances of success, particularly if we can focus on a few key regions for more intensive investigation.

References

Hinze, L.L. 2003. Molecular marker analysis of population genetic structure and progress from reciprocal recurrent selection in two Iowa maize (*Zea mays* L.) populations. PhD dissertation.

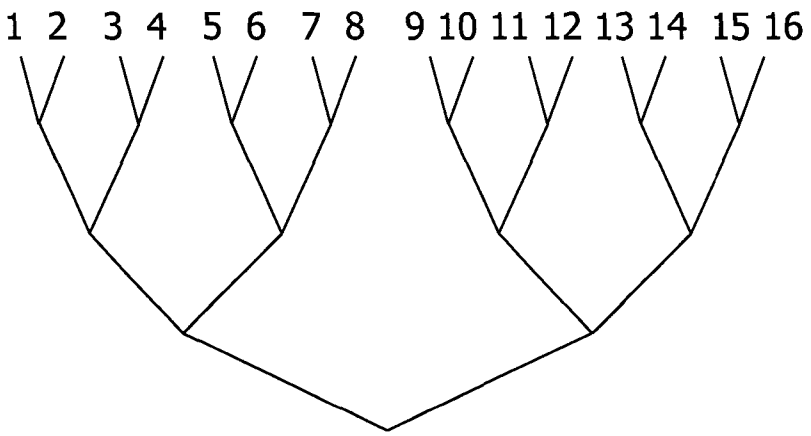
Labate, J.A., K.R. Lamkey, M. Lee, and W.L. Woodman. 1999. Temporal changes in allele frequencies in two reciprocally selected maize populations. Theor. Appl. Genet. 99:1166-1178.

Lynch, M. 1996. A quantitative-genetic perspective on conservation issues. *In* J.C. Avise and J.L. Hamrick (ed.) Conservation genetics: Case histories from nature. Chapman and Hall, New York.

APPENDIX: SUPPLEMENTAL TABLES AND FIGURES

Table 1. Inbred progenitors and mating scheme to form Iowa Stiff Stalk Synthetic (BSSS).

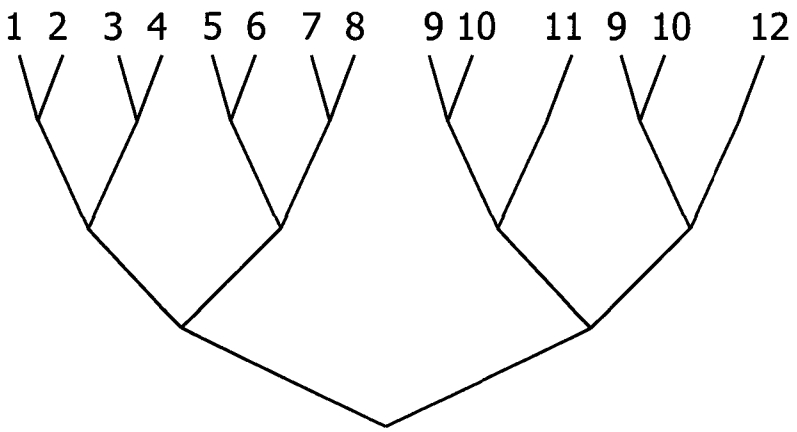
#	Inbred	#	Inbred
1	I159	9	Ill Hy
2	I224	10	Oh3167B
3	Os420	11	Ind AH83
4	WD456	12	Tr 9-1-1-6
5	Ind 461-3	13	F1B1
6	Ill 12E	14	A3G-3-1-3
7	CI617	15	CI187-2
8	CI540	16	Le23



BSSS Cycle 0

Table 2. Inbred progenitors and mating scheme to form Iowa Corn Borer Synthetic #1 (BSCB1).

#	Inbred	#	Inbred
1	R4	9	P8
2	K230	10	L317
3	I205	11	CC5
4	Oh40B	12	A340
5	Ill Hy		
6	Oh07		
7	Oh33		
8	Oh51A		



BSCB1 Cycle 0

Table 3. Markers used in this study. Markers that were omitted from the analyses (listed near the bottom of the table) were dropped due to excess heterozygosity, excess missing data, monomorphism, or ambiguous results.

Marker	Chrom.	Bin	Repeat Sequence	Maize GDB address (http://www.maizegdb.org /cgi-bin/displaylocusrecord.cgi?id=...)
bnlg1012	09	9.04	AG	144764
bnlg1036	02	2.06	AG	144778
bnlg1064	02	2.03	AG	144794
bnlg1175	02	2.04	AG	144833
bnlg1233	02	2.08	AG	144850
bnlg1297	02	2.02	AG	144871
bnlg1316	02	2.08	AG	144875
bnlg1521	06	6.07	CT	144923
bnlg1523	03	3.03	AG	144924
bnlg1746	02	2.08	AG	144985
bnlg1839	10	10.07	AG	145012
bnlg1893	02	2.09	AG	145023
bnlg2190	10	10.06	AG	145060
mmc0381	02	2.08	GA	239583
phi048	05	5.07	ATCG	41372
phi082	07	7.05	AG	64580
phi123	06	6.07	AAAG	113913
phi126	06	6.00	AG	111774
phi129	06	6.05	ATAC	111775
phi333597	05	5.05	AAG	256162
umc1002	06	6.00	TA	167610
umc1010	03	3.09	GA	120441
umc1024	02	2.04	GA	167283
umc1049	02	2.08	CT	231182
umc1054	10	10.04	CAG	194081
umc1058	04	4.11	GC	200389
umc1062	03	3.09	AGC	199156
umc1066	07	7.01	GCCAGA	12495
umc1067	04	4.04	GCC	40975
umc1075	08	8.01	ATTGC	200367
umc1078	09	9.05	GT	207068
umc1106	01	1.00	GAGA	235042
umc1107	09	9.04	TC	235045
umc1120	09	9.04	GGCAT	235084
umc1121	08	8.05	AGAT	235087
umc1123	01	1.06	AC	235093
umc1125	07	7.04	CTCG	235099
umc1134	07	7.03	AGC	235126

Table 3. Continued.

Marker	Chrom.	Bin	Repeat Sequence	Maize GDB address (http://www.maizegdb.org /cgi-bin/displaylocusrecord.cgi?id=...)
umc1135	03	3.07	TCA	235129
umc1136	03	3.09	GCA	235132
umc1139	08	8.01	GAC	235141
umc1141	08	8.06	GA	235147
umc1147	01	1.07	CA	235165
umc1153	05	5.09	TCA	235183
umc1154	07	7.05	AC	235186
umc1156	02	2.06	AAT	235192
umc1165	02	2.01	TA	235219
umc1166	01	1.02	CT	235222
umc1170	09	9.02	TC	235234
umc1171	05	5.04	GTT	113958
umc1172	08	8.04	CCA	25423
umc1178	06	6.02	GGC	143267
umc1185	02	2.03	GC	65609
umc1187	06	6.05	CCT	51346
umc1196	10	10.07	CACACG	235312
umc1202	08	8.05	GGC	12064
umc1227	02	2.01	AGG	242203
umc1229	06	6.01	AG	242206
umc1231	09	9.05	GA	248528
umc1232	04	4.00	ACAG	242212
umc1236	08	8.03	TGCA	242224
umc1239	10	10.03	TG	242233
umc1241	07	7.00	GTCTTTG	242239
umc1246	10	10.04	AAAT	242254
umc1259	02	2.04	GCG	246207
umc1262	02	2.02	GTC	246216
umc1270	07	7.01	GCA	246240
umc1271	09	9.03	CGG	246243
umc1272	10	10.04	CTAGC	246246
umc1273	03	3.08	AAG	246249
umc1274	05	5.03	TGC	246252
umc1276	04	4.01	GGC	246258
umc1284	04	4.09	TCA	246282
umc1286	03	3.07	TCCT	246288
umc1289	08	8.03	TCG	246297
umc1290	01	1.09	GCG	246300
umc1292	01	1.01	TGG	246306

Table 3. Continued.

Marker	Chrom.	Bin	Repeat Sequence	Maize GDB address (http://www.maizegdb.org /cgi-bin/displaylocusrecord.cgi?id=...)
umc1301	07	7.03	GCA	246333
umc1303	04	4.05	CCG	246339
umc1304	08	8.02	TCGA	246342
umc1309	08	8.05	TG	25708
umc1311	03	3.06	TCTT	248531
umc1316	08	8.05	GA	248546
umc1320	03	3.08	GAAC	248558
umc1324	07	7.03	AGC	248570
umc1326	02	2.04	GCC	248576
umc1328	04	4.09	TGC	248582
umc1330	10	10.04	GCG	248588
umc1366	09	9.06	TCC	248696
umc1367	10	10.03	CGA	248699
umc1370	09	9.01	CGGG	248708
umc1381	10	10.03	AAC	256252
umc1382	04	4.05	AAC	256255
umc1388	06	6.05	CGC	256273
umc1390	04	4.05	CCT	256279
umc1393	07	7.02	GTC	256288
umc1403	01	1.03	GCA	256318
umc1406	07	7.05	CTCA	256324
umc1409	07	7.01	GCTC	256333
umc1411	01	1.09	TGG	256339
umc1412	07	7.04	CCA	256342
umc1414	08	8.01	GCTA	256348
umc1416	05	5.00	CAA	256354
umc1421	01	1.11	ACAA	256369
umc1424	06	6.06	TCC	256378
umc1431	01	1.09	GCA	256399
umc1444	06	6.01	CAC	256408
umc1446	01	1.08	TAA	256414
umc1448	02	2.04	GCT	256420
umc1452	01	1.03	GCC	256432
umc1453	10	10.04	GCG	256435
umc1457	08	8.03	GTG	256447
umc1458	03	3.02	GCT	256450
umc1465	02	2.04	ACACA	256471
umc1467	01	1.02	CTT	256477
umc1470	08	8.03	TAA	256486

Table 3. Continued.

Marker	Chrom.	Bin	Repeat Sequence	Maize GDB address (http://www.maizegdb.org /cgi-bin/displaylocusrecord.cgi?id=...)
umc1471	08	8.03	TGG	256489
umc1479	01	1.03	AGA	256513
umc1483	08	8.01	ACG	256525
umc1489	03	3.07	GCG	272887
umc1492	09	9.04	GCT	272896
umc1496	05	5.00	GCA	272908
umc1498	06	6.01	CAG	272914
umc1502	05	5.05	CA	272926
umc1506	10	10.05	AACA	272938
umc1508	01	1.06	ATG	272944
umc1509	04	4.02	TG	272947
umc1514	01	1.03	AGT	272962
umc1525	02	2.09	CGA	272995
umc1528	03	3.07	TGCG	273004
umc1539	03	3.05	GGC	273037
umc1545	07	7.00	AAGA	273058
umc1550	04	4.03	CT	82183
umc1559	04	4.08	ATG	273097
umc1567	07	7.03	AGA	273121
umc1568	01	1.02	TAG	273124
umc1576	10	10.02	TG	114010
umc1580	02	2.04	CCG	273160
umc1588	09	9.01	AT	273184
umc1596	09	9.01	GGC	273208
umc1603	01	1.05	GCC	273229
umc1604	02	2.08	GCC	273232
umc1624	05	5.04	CAG	275140
umc1632	07	7.01	AGC	291228
umc1634	09	9.03	AG	291234
umc1639	03	3.09	TGTCC	291249
umc1641	03	3.09	TCGCC	291255
umc1644	03	3.06	TTG	291264
umc1648	10	10.04	TC	291276
umc1652	04	4.04	CTGGA	291288
umc1657	09	9.05	GACGG	291303
umc1663	08	8.09	ATG	291321
umc1671	07	7.05	AGC	291345
umc1675	09	9.07	CGCC	291357
umc1678	10	10.04	TCG	291366

Table 3. Continued.

Marker	Chrom.	Bin	Repeat Sequence	Maize GDB address (http://www.maizegdb.org /cgi-bin/displaylocusrecord.cgi?id=...)
umc1682	04	4.01	AC	40807
umc1684	07	7.03	CGC	292398
umc1688	09	9.03	GGA	292410
umc1690	03	3.07	GCA	292416
umc1691	09	9.03	AG	292419
umc1698	09	9.02	GTA	292440
umc1700	09	9.03	TAG	292446
umc1706	01	1.07	TCG	292464
umc1707	04	4.11	AT	292467
umc1719	04	4.10	GCG	292503
umc1724	08	8.06	CGA	292518
umc1737	01	1.11	AGA	292557
umc1739	10	10.03	ATAC	292563
umc1743	09	9.03	GGC	292575
umc1746	03	3.01	CAC	292584
umc1747	05	5.04	CTG	292587
umc1752	05	5.06	CGG	292602
umc1760	07	7.05	GA	292626
umc1766	05	5.01	CGCCGG	292644
umc1778	08	8.03	GTC	292680
umc1779	06	6.07	TCG	292683
umc1780	03	3.01	ACC	292686
umc1788	07	7.00	AAAAT	292710
umc1792	05	5.08	CGG	292722
umc1793	03	3.00	AT	292725
umc1812	01	1.03	ACC	301476
umc1813	03	3.09	CAG	301479
umc1814	03	3.02	CGA	301482
umc1819	01	1.12	CAAC	301497
umc1821	04	4.03	AGT	301503
umc1829	05	5.09	AG	301527
umc1844	03	3.08	TC	301572
umc1847	04	4.07	CGC	301581
umc1853	05	5.05	GT	301599
umc1857	06	6.04	TAA	301611
umc1869	04	4.06	GGT	301647
umc1875	02	2.06	CT	301665
umc1887	06	6.03	CGA	301701
umc1912	06	6.06	GCG	301776

Table 3. Continued.

Marker	Chrom.	Bin	Repeat Sequence	Maize GDB address (http://www.maizegdb.org /cgi-bin/displaylocusrecord.cgi?id=...)
umc1930	10	10.04	GT	301830
umc1934	02	2.02	AT	301842
umc1950	08	8.05	na	301890
umc1959	08	8.05	na	301917
umc1967	09	9.01	na	301941
umc1970	03	3.01	na	301950
umc1979	06	6.04	na	301977
umc1980	02	2.02	na	301980
umc1983	07	7.02	na	301989
umc1984	08	8.03	na	301992
umc1986	07	7.02	na	301998
umc1994	04	4.07	na	302022
umc1996	06	6.00	na	302028
umc1997	08	8.06	na	415181
umc2007	02	2.04	na	302061
umc2016	10	10.03	ACAT	309023
umc2017	10	10.03	CAA	309026
umc2021	10	10.07	TGG	309038
umc2025	01	1.05	AGCT	309050
umc2027	04	4.06	AAAG	309056
umc2030	02	2.04	CGA	309065
umc2031	08	8.06	GCG	309068
umc2037	08	8.06	CTGT	309086
umc2040	06	6.05	CGC	309095
umc2043	10	10.05	TCC	309104
umc2053	10	10.01	CGA	309134
umc2056	06	6.01	ATC	309143
umc2057	07	7.02	GCT	309146
umc2059	06	6.08	CAG	309152
umc2067	10	10.03	CATG	309176
umc2173	08	8.03	CGT	485112
umc2182	08	8.04	TCC	485139
umc2190	07	7.06	CCT	485163
The markers below were dropped from the analyses for the various reasons described in the caption.				
bnlg1144	03	3.02	na	144822
bnlg1327	02	2.02	CT	144878
bnlg1633	02	2.07	AG	144954
bnlg1810	09	9.01	AG	145002

Table 3. Continued.

Marker	Chrom.	Bin	Repeat Sequence	Maize GDB address (http://www.maizegdb.org /cgi-bin/displaylocusrecord.cgi?id=...)
mmc0092	01	1.025	CGG	167129
phi041	10	10.00	CGG	111816
phi052	10	10.02	AAG	40923
phi438301	04	4.05	ACC	256207
umc1004	02	2.06	CA	167235
umc1115	10	10.04	AG	235069
umc1122	01	1.06	CGT	235090
umc1128	01	1.07	TC	235108
umc1180	04	4.10	CATG	235264
umc1242	07	7.05	TAA	242242
umc1253	05	5.00	TTC	242275
umc1299	04	4.06	AAG	246327
umc1350	06	6.07	GCT	248648
umc1355	05	5.03	TTTC	248663
umc1362	04	4.05	AGT	248684
umc1375	05	5.07	CCG	256234
umc1383	01	1.08	GACG	256258
umc1399	03	3.07	CTAG	256306
umc1419	02	2.00	AGT	256363
umc1422	02	2.02	GCC	256372
umc1426	07	7.00	AGAGG	256384
umc1472	01	1.04	TTC	256492
umc1482	05	5.05	AGC	256522
umc1491	05	5.00	AGA	272893
umc1500	01	1.11	ACC	272920
umc1516	02	2.08	TA	272968
umc1542	02	2.02	AG	273046
umc1571	09	9.04	CAA	275131
umc1578	03	3.09	GCG	273154
umc1590	01	1.06	AAGGAG	273190
umc1613	01	1.00	TCG	273259
umc1617	08	8.03	ATT	273271
umc1636	09	9.02	ACTGC	291240
umc1646	05	5.07	CTGGA	291270
umc1660	07	7.03	ACG	291312
umc1677	10	10.05	GGC	291363
umc1695	07	7.00	CA	292431
umc1704	02	2.08	AGG	292458
umc1733	09	9.06	CATC	292897

Table 3. Continued.

Marker	Chrom.	Bin	Repeat Sequence	Maize GDB address (http://www.maizegdb.org /cgi-bin/displaylocusrecord.cgi?id=...)
umc1757	04	4.01	TCC	292617
umc1786	08	8.01	TC	292704
umc1789	09	9.06	CCG	292713
umc1818	06	6.02	CAG	301494
umc1861	02	2.04	AT	301623
umc1915	03	3.08	ACA	301785
umc1957	09	9.00	na	301911
umc1960	08	8.06	na	301920
umc1962	10	10.03	na	301926
umc1972	01	1.06	na	301956
umc2018	10	10.01	CCT	309029
umc2044	04	4.10	CGG	309107
umc2052	08	8.08	GGA	309131
umc2069	10	10.02	GCG	309182

Figure 1. Distribution of gene diversity from Monte Carlo simulations sampling eight inbred lines from the BSCB1 progenitor groups. Blue bars represent simulated gene diversity outcomes. Yellow line is the normalized distribution of the simulated gene diversity. Yellow vertical bar represents the observed gene diversity within the CBL (there is no importance to the height of this bar—only its *x*-axis position).

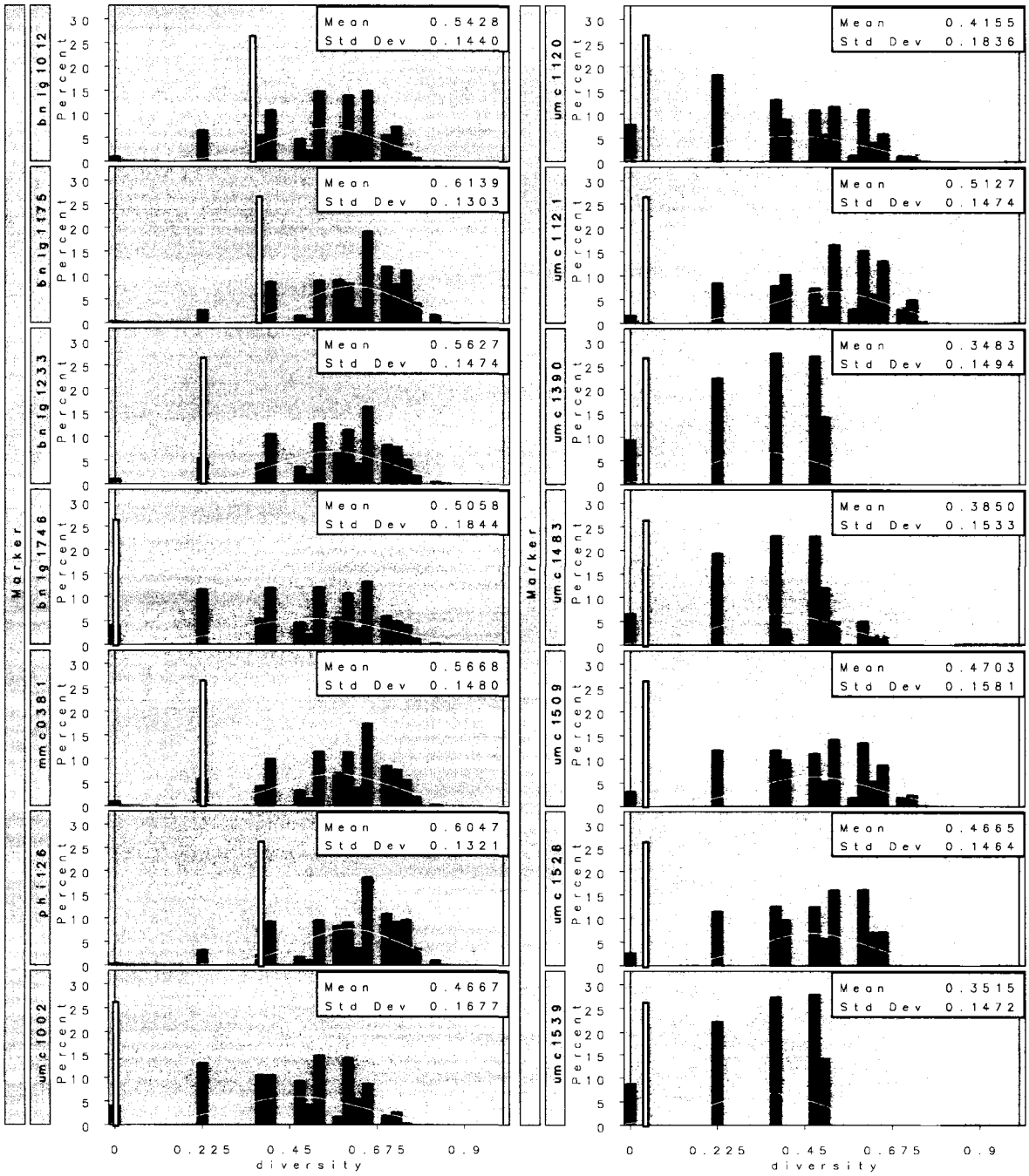


Figure 1. Continued.

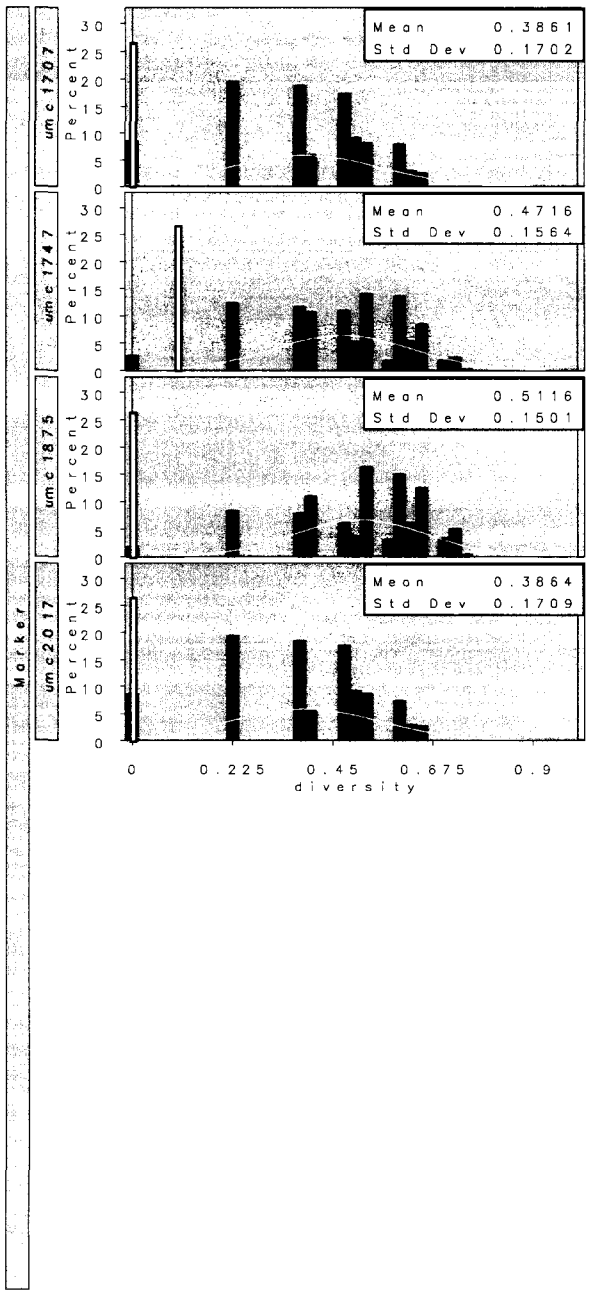


Figure 2. Distribution of gene diversity from Monte Carlo simulations sampling 14 inbred lines from the BSSS progenitor groups. Blue bars represent simulated gene diversity outcomes. Yellow line is the normalized distribution of the simulated gene diversity. Yellow vertical bar represents the observed gene diversity within the SSL (there is no importance to the height of this bar—only its *x*-axis position).

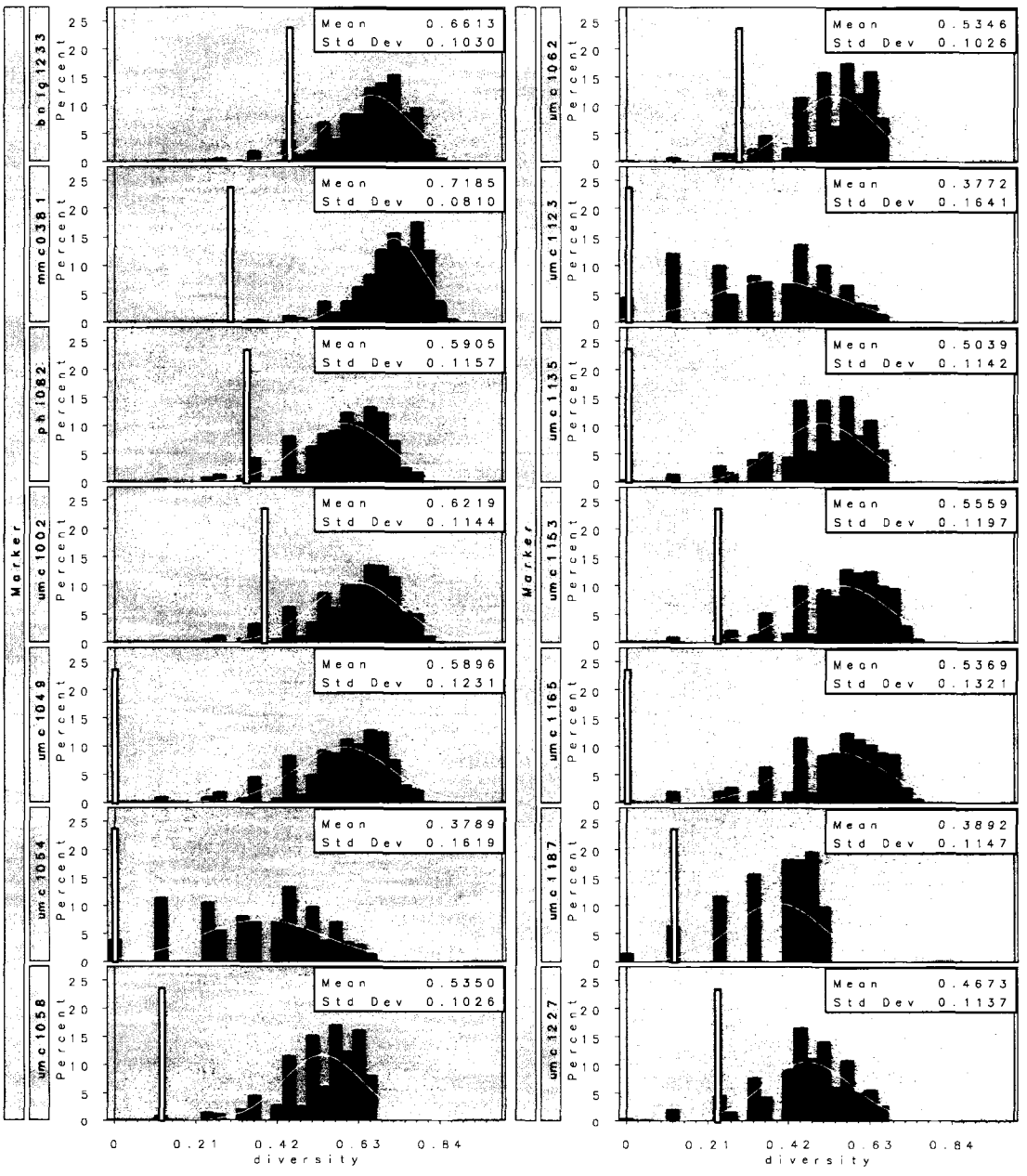


Figure 2. Continued.

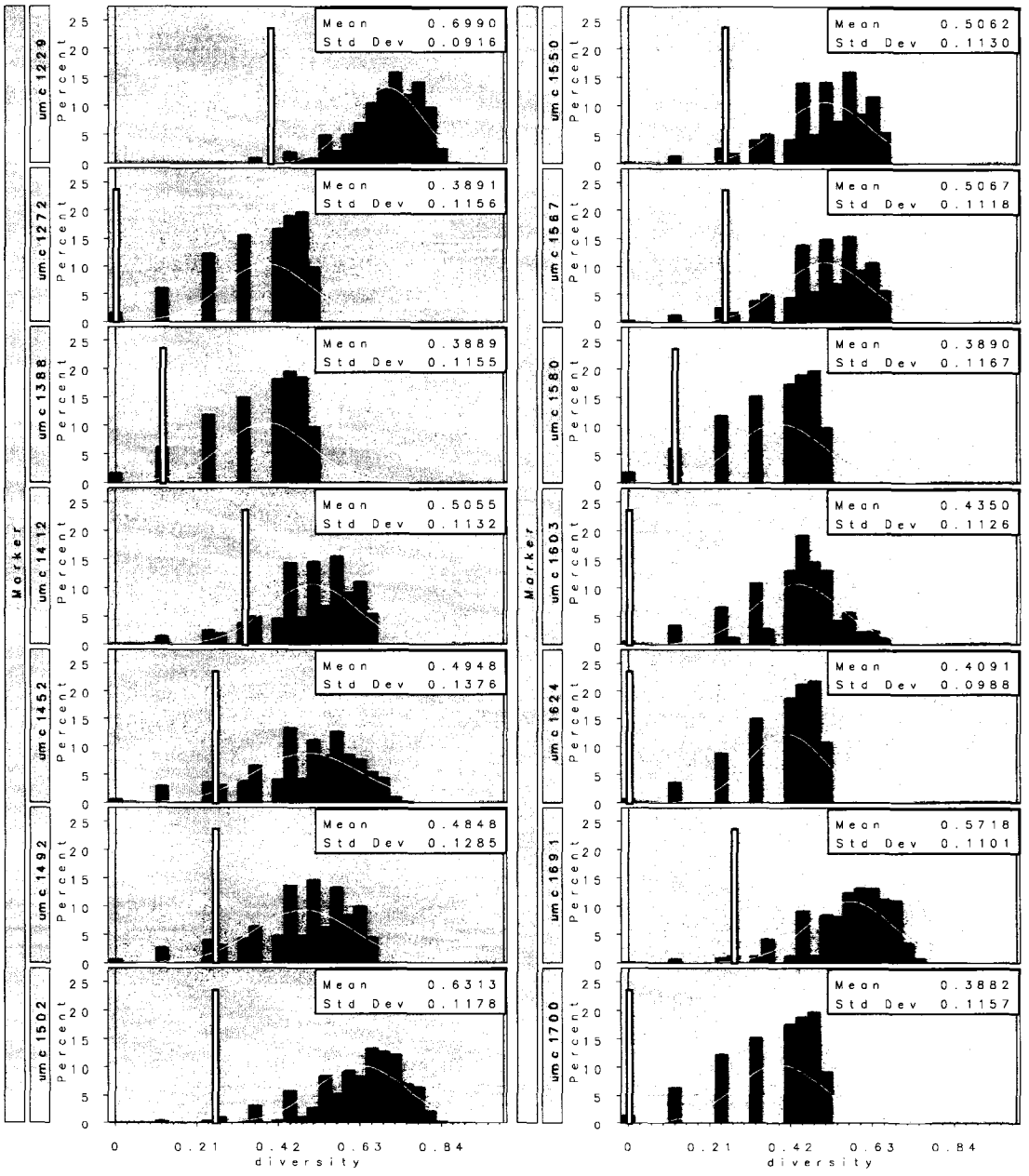
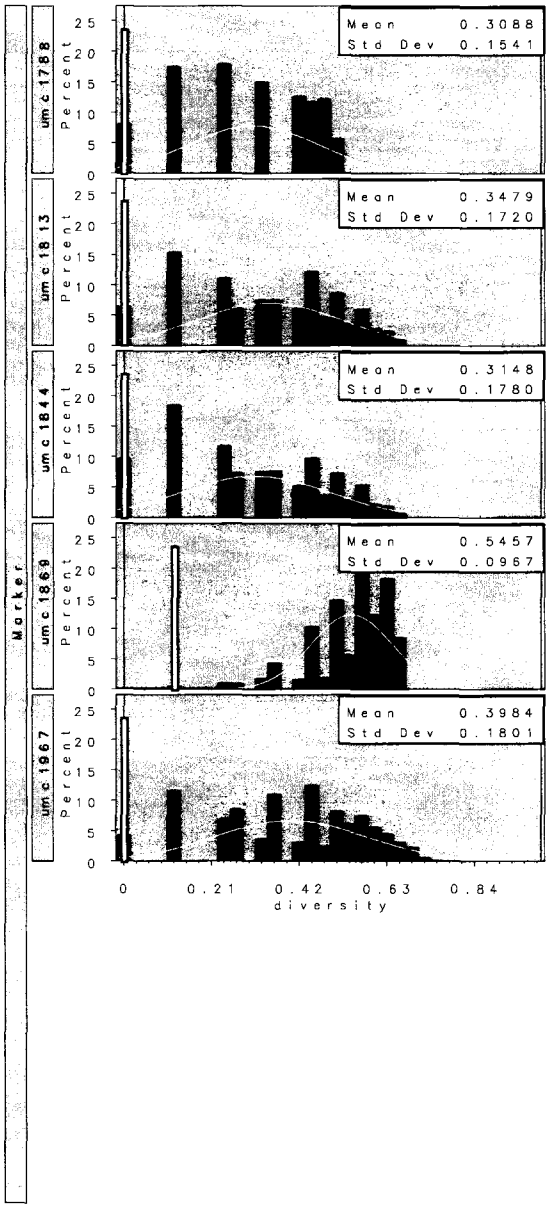


Figure 2. Continued.



ACKNOWLEDGEMENTS

Glenda is the most important person in my life, for so many reasons that don't matter to anybody but us. Next in line comes family—Leda, Jess & Chad, parents, siblings, and parents-in-law. But nobody's going to read this anyway, so I have a chance to spout off about a lot of things that have influenced me in a (mostly) positive manner during the last few years. In no particular order, and for a wide range of reasons, I wish to express varying amounts of gratitude to each of the following: Glenda, for everything, including putting up with me; Leda, for making me proud, and for putting up with me; Jess, for making me proud and letting me be part of her life, no matter how hard I fought it—and for putting up with me; Chad, for being Chad; Demeter; Lori Lynn Hinze; KRL; Jode; BMW; Ceres; Clint 'Creech' Turnbull, for living a life that demonstrates some things are more important than others; Bacchus; Dogpile & Google; Glenn ("What the Christ?"); Ma & Pa, for teaching me the right stuff, even when I practiced the wrong stuff; Arnel Hallauer; Nason, Brummer, Jean-Luc & Mike Lee; Cap'n Dale, for free tuition at DTU; Jaci & Mary; Billy Idol, Alice Cooper, and AC/DC; Sisyphus; Peter Peterson; my friends at the Knob Creek Network, for inspiration; Harvey Byrd, for cheering me on; Betty Byrd, for worrying about me; chairs with just-the-right butt groove; the Beer Rats, for turning out much better than any parents ever imagined possible (I don't know why all of you look 20 years older, I haven't changed a bit); Dob, for proving that you don't need hair to be a prrrtttty man; Ron, for his excellent choice of parents, brother, and wife; Dano, mainly because he still looks like Roger Daltry; Pierce, for being the oldest little kid I know; sis Julie, for the 3 haircuts in the last 3 years, and the 4 more haircuts in the last 2 months; Dionysus; Bystrak, because he's better than us, he just

doesn't act like it; Fernando, for the GS strategy ("it's just a game of survival"); GT & the DDs; Uncle Teddy; Eddie, Alex, Michael, and the Red Rocker; Bidlack, for being a goofball; high ceilings, low floors; Frodo, Tater, Fargo, & Lucy ("ju got som splaining to doo!"); Centeotl and Xilonen; steak; Dublin Bay; George Dickel; so many people for being exceptional role models in many different areas; some of the same people, and most of the rest, for providing excellent examples of how NOT to do things.

If you're reading this, and you think you should have been included, you were. I just got sleepy before completing this page. But I was thinking about you, too.